RICE UNIVERSITY

Improving Protein Conformational Sampling by Using Guiding Projections

by

Anastasia Novinskaya

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Science

Approved, Thesis Committee:

Dr. Lydia E. Kavraki, Chair
Noah Harding Professor
Department of Computer Science

Dr. Luay Nakhleh
Associate Professor
Department of Computer Science

Dr. Christopher M. Jermaine
Associate Professor
Department of Computer Science

Houston, Texas

December, 2015

ABSTRACT

Improving Protein Conformational Sampling by Using Guiding Projections

by

Anastasia Novinskaya

The ability of a protein to perform its function is mainly defined by the spatial shape it exists in and the way the protein alternates between several stable shapes. To prevent or cure diseases related to protein malfunctioning we study the conformational space of proteins. Sampling-based motion planning algorithms from the field of robotics have been very successful at this task. However, studying the conformational space of large proteins with hundreds or thousands of Degrees of Freedom remains a big challenge. In this work we investigate how the dimensionality curse can be mitigated by means of low-dimensional projections. Our experiments demonstrate that incorporating the information available on the studied protein into the projection can benefit the conformational exploration process. The techniques we developed to generate efficient low-dimensional projections can enable sampling-based planners to study protein systems, such as viruses, that are currently too large to be investigated by other methods.

Acknowledgment

I would like to express my deepest gratitude to my adviser, Dr. Lydia Kavraki, who has supported, motivated, and guided me throughout the course of this work. With her own example she always inspires me to strive to deliver greater outcomes. I also want to thank her for providing her students with an amazing environment for research and collaborations.

I am also very grateful to Dr. Mark Moll and Dr. Didier Devaurs for their tremendous help and support in the current project. Our regular discussions always inspire me with the scientific curiosity and help me to stay on track with my research goals.

Also I would like to thank all members of KavrakiLab. It is a big honor for me to be a part of this group, and I truly feel like a part of big family with these people.

Additionally, I would like to thank members of my thesis committee, Dr. Luay Nakhleh and Dr. Christopher M. Jermaine, for their time and the feedback they provided for this thesis.

Finally, I want to express my deepest gratitude to my beloved parents. I especially thank my grandmother, who is my constant source of inspiration for growth and self-development.

Contents

Chapter 1

Introduction

Proteins are involved in almost every process within living organisms. Often a protein's activity is modulated/characterized by its ability to switch among several stable conformations. Understanding how a protein shifts between these states is essential for treating or preventing diseases related to the protein's dysfunction [1]. However, the shift happens so rapidly that it is extremely hard to monitor it experimentally. For this reason, a common approach to gain knowledge of how proteins move is to model this process computationally. There exist several classes of algorithms for simulating protein motion. They vary from highly physically precise and computationally expensive simulation techniques, such as Molecular Dynamics (MD) [2], to methods producing fast but rather approximate analysis of protein motions, such as Normal Mode Analysis (NMA) [3] and Elastic Network Models (ENM) [4]. The development of new advanced technologies (e.g., serial femtosecond crystallography [5], small-angle X-ray scattering [6], and advanced NMR methods [7], among others) enables us to extract a protein's structure at high resolution and even capture a protein in different conformations. However, such technologies are usually very expensive and time-consuming; furthermore, they can only capture snapshots of a protein in a particular moment, and not its motion.

1.1 Sampling-based Search Algorithms

Our work on modeling protein flexibility involves sampling-based motion planning algorithms that have been adapted from the field of robotics. These methods fill the gap between the two classes of approaches mentioned above: they represent a trade-off between physical accuracy and computational cost. As a result, these methods have been very efficient at producing representative large-scale protein motions [8, 9]. Sampling-based algorithms explore the conformational space of a protein by randomly sampling it (usually using a special heuristic) and constructing a graph where each node represents a feasible low-energy protein conformation (or state), and each edge represents a possible low-energy local transition between two states. The computed graph describes the topology of a protein's energy landscape and the connectivity of its low-energy areas. This graph can be used to find possible large-scale transitions between two given protein conformations.

1.2 Expansive Search

All current computational methods for modeling protein flexibility, including sampling-based techniques, suffer from the curse of dimensionality: their complexity grows exponentially with an increasing number of dimensions. Moreover, large proteins are also highly-constrained systems, which increases simulation complexity even further. To mitigate these problems the specific, so-called "expansive," approach has been developed. Expansive planners present a current state-of-the-art way to model

middle-sized or large proteins using the sampling-based method. These planners grow their conformational graph by iteratively applying an expansion procedure. More precisely, such planners use the conformations which already belong to their graph and therefore are valid low-energy protein structures to generate a new conformation by a slight perturbation. This approach is far more effective for the considered problem than other sampling-based methods commonly used in robotics applications (such as Rapidly-exploring Random Trees (RRT) planners, for example). Non-expansive planners often employ absolutely random configuration of a system to generate a new state. However, in case of modeling proteins, a random state almost always represents an invalid high-energy protein conformation, which could not be used to grow a conformational tree.

In the current work we focus on two specific expansive planners: Expansive Space Tree (EST) [10] and Kinematic Planning by Interior-Exterior Cell Exploration (KPIECE) [11]. These planners use linear projections to assess how they cover the conformational space of a protein with the samples they produce. Based on this information, they direct their search towards unexplored regions of the conformational space.

1.3 Low-dimensional Projections

Even though large proteins have thousands of Degrees of Freedom (DoFs), the extensive analysis of protein conformations generated by various methods (such as MD [12],

X-Ray Crystallography [13], or Normal Mode Analysis [14]) has shown that the majority of their residues move in a correlated fashion. As a result, protein motions can usually be characterized by just a few collective DoFs [15, 16]. Therefore, projections that are aligned with the low-dimensional manifold of protein motion can represent a good approximation of the high-dimensional conformational space of a protein. There exist variety of different methods to construct such low-dimensional projections. These methods could be divided into two main categories: non-linear and linear. Proteins are believed to move in a highly non-linear fashion. Therefore, non-linear methods usually present more precise approximation of the protein conformational space. However, non-linear methods introduce a large computational overhead on the projection procedure. For this reason, in this work we only examine the performance of different low-dimensional *linear* projections. For expansive planners, this kind of low-dimensional projection constitutes the main instrument to handle dimensionality and, thus, plays an essential role in their success. Even so, there exists no general method for constructing a good projection - a projection that enhances the conformational exploration in terms of runtime and/or volume of the explored conformational space. Traditionally, projections for expansive planners are generated randomly. In this thesis we address the questions of whether the random projections are good enough for modeling proteins and whether there are methods to generate better projections.

1.4 Contributions

In this work we first introduce a new methodology to construct effective low-dimensional projections using simple biological knowledge available for a given protein. We demonstrate that such "expert" projections can improve the process of conformational search for expansive planners compared to the projections of the similar dimensionality generated randomly. Then we generalize the described methodology into the algorithm that will generate a successful projection automatically. We have applied the constructed projections to two different kinds of conformational search problems: 1) finding a feasible low-energy transition between two given protein states, and 2) exploring the conformational space starting from a given protein state. In the case of the first problem, the expert and automatically generated projections show improvements in algorithm runtime, and, in the case of the second problem, improvements in space coverage. In the final part of this work we go beyond the investigation of what a good linear projection is and question the general belief that low-dimensional projections are *always* useful for conformational exploration.

1.5 Organization

The rest of this thesis is organized as follows: Chapter 2 presents the related work. Chapter 3 describes the context of our work: two problems of the conformational search that we use to benchmark the proposed projection methods; the framework that we use to perform a conformational search; and two projection-based expan-

sion heuristics that we consider in this work. Chapter 4 introduces the methods for constructing different types of projections. This chapter also describes the four middle-sized protein systems (having at least one hundred residues) we tested using our approach: Cyanovirin-N, Calmodulin, Ribose-binding protein, and Adenylate Kinase, and the associated with them projections. Chapters 5 and 6 report the performance comparison of the random projections with the expert and auto projections respectively. Chapter 7 questions the state-of-the-art concept of using a projection to guide the exploration, and investigates the cases when the application of a linear projection is actually beneficial. Finally, Chapter 8 concludes the thesis and presents some of our future work.

Chapter 2

Related Work

2.1 Sampling-based Path Planning Methods

In recent years sampling-based path planning methods have been very successful in the field of robotics. These techniques address the problem of finding a path from a start point to a goal point for a studied system in a given environment. The progress achieved in the past few decades has made sampling-based methods applicable to systems with a large number of DoFs, such as proteins. In fact, proteins can be defined as high-dimensional systems of links and joints. Their motions can be modeled by sampling-based algorithms in the same way as motions of articulated chains in robotics [9].

Sampling-based methods have been very effective for the fast computation of representative motions of molecular systems [8]. A broad range of approaches exploit sampling-based techniques to address various biological problems, such as exploring energy landscapes [17], modeling protein folding pathways [18], analyzing protein loops [19], or modeling large-scale transitions in a protein structure [20].

The sampling-based methods explore the conformation space (i.e., the space of all possible combinations of values that the system's DoFs can take) of a system and

build a graph connecting the feasible conformations. At each step, a sampling-based algorithm samples a conformation. Then it performs a validity check for the chosen sample. In protein modeling, this means eliminating high-energy protein conformations. If the sampled state satisfies all the constraints of the problem, it is added to the graph as a new node, otherwise it is discarded. Finally, the valid states are connected into a graph structure by adding edges between the nearest configurations. Edges also often undergo a validity test: only the edges that satisfy the system's constraints are added to the graph. The constructed graph represents the topology of conformation space: the nodes represent the low-energy clash-free conformations of the protein, and the edges represent feasible local transitions between the corresponding conformations.

2.2 Using Projections to Guide Conformational Sampling

Despite the capability of sampling-based methods to generate large-scale protein motions much faster than physics-based simulations, they still suffer from the curse of dimensionality. Middle-sized and large proteins require hundreds or thousands of variables to encode a conformation. Moreover, because large proteins often represent highly-constrained systems, they can only move in a very limited fashion. These issues represent a significant challenge for sampling-based approaches as their complexity grows exponentially with the dimensionality of the system as well as with the decrease in volume of the space of low-energy conformations.

In this work we use a group of expansive planners that were developed to specifically tackle high-dimensional and highly-constrained problems [10, 11]. Expansive planners iteratively grow a tree of feasible protein conformations by choosing a state which is already in the tree (and therefore has low energy), and slightly perturbing some DoFs of that state to generate a new, child conformation. These planners employ a low-dimensional projection of a protein conformational space to store statistics of the exploration progress. To identify a promising parent state for further expansion, these planners use the estimate of conformational space coverage provided by the projection.

Various approaches have been developed in the context of sampling-based methods in general and "expansive" algorithms in particular to overcome the curse of dimensionality. One of the common techniques is to identify flexible and rigid parts of a protein on-the-fly and focus computations on exploring mainly the flexible regions. The framework employed in the current work has the functionality of identifying the flexible protein regions automatically based on a protein's secondary structure. An alternative approach for rigidity analysis based on the pebble game computations [21] is presented in the works of Amato's group (see, for example, [22]), and Streinu's group [23]. The rigidity analysis technique has been mainly used to bias the local search (choosing the local protein motion) [22, 24]. In the current work, we focus on enhancing the global search (choosing the conformation from the tree for further expansion) by estimating the effect of a low-dimensional linear projection on the overall

exploration process.

Many recent approaches use the expansive planners to explore protein conformational landscapes [25–27]. They suggest a variety of algorithms for computing low-dimensional projections, including: simplistic 1D projections based on lRMSD towards a goal structure (or some milestone) [27]; slightly more advanced 3D projections computed from mean interatomic distances to the given points of the structure [25]; and the quite intricate 1D projections generated from the contact matrix with usage of hashing algorithms [26]. Often, the aforementioned projections are also combined with 1D projection layer based on the energy of the structure [26]. All of these methods have some biological intuition to support them. However, there exists very little analysis of how the chosen projection methods affect the conformational exploration. From the above-mentioned works only [26] provides the comparison between performance of two different projection algorithms: 1) ultrafast shape recognition (USR) [28], and 2) hashing based on structural profiles. USR method in [25] maps the protein conformation onto a 3D coordinate space by computing mean atomic distances from the centroid atom (ctd) of a protein, from the atom that is the farthest from the centroid (fct), and from the atom that is the farthest from fct (ftf). The authors claim that these three coordinates are able to capture the overall spatial organization of a protein. They apply the described approach as a guiding mechanism for conformational search investigating the structural characteristics of protein native states. In their later work, [26], Shehu and Olson introduce a novel guiding projection

mechanism and show its improvement over the simplistic USR method. The second guiding approach is based on the protein structural profiles [29]. For a protein with N residues, the structural profile represents a NxN matrix, where the element m_{ij} is set to 1 if the residues i and j are in the contact (i.e., they are not close in the protein sequence, $|i - j| > 3$, but they exist in the spatial proximity, $|C_{\alpha i} - C_{\alpha j}| < d$, where d is a parameter of the method), and otherwise it is set to 0. The main eigenvector (principle component) or the weighted linear combination of all principle components of this matrix represents a vector containing the structural information about the protein. This vector is then used to generate a hash function, which plays the role of a 1D projection. The authors show in [26] that the described algorithm produces better conformational exploration than USR coordinates. However, the procedure of computing the hash function based on the structural profiles is much more expensive than USR projection or linear projections, which we study in the current work.

In this thesis we investigate the role of the linear projections on the process of conformational search. Prior work [30] studies the influence of such projections in the context of robotic systems with at most a few dozen of DoFs. That work demonstrates that some projections enhance sampling-based planners more than others, even for systems with moderate dimensionality. Because the conformational space grows exponentially with the number of dimensions, the importance of proper guidance increases significantly for high-dimensional systems. For this reason we believe that for high-dimensional systems the difference between "successful" and "unsuc-

cessful" projections is even more drastic. In [30] the authors find that the projection showing the best performance usually belongs to the group of randomly generated projections. Currently, in many cases, low-dimensional projections for sampling-based planners are chosen randomly. However, the described conclusions cannot be applied to protein modeling without additional investigation. Proteins represent significantly larger systems than the ones considered in the above-mentioned paper. When the dimensionality of a system increases, chances of constructing a "good" low-dimensional projection randomly diminish greatly. Furthermore, the user-defined projections in the analyzed paper are built under the assumption that a projection is independent from the environment of the system. In the case of proteins, the environment is encoded by their energy landscape: it defines which parts of the protein are mostly rigid and which parts could change their shape and participate in large-scale conformational transitions. This information provides essential insight for enhancing conformational search. If an expert projection is tailored for the efficient exploration of a particular protein's energy landscape, the same projection will not benefit the investigation of another protein.

In our work, we evaluate the performance of expert-defined projections that incorporate any available information about a protein's flexibility. We demonstrate that our expert projections accelerate and enhance the exploration of the conformational space compared to the traditionally-used, randomly-generated projections. We define an algorithm to construct an effective projection automatically. In the final part

of this thesis we investigate whether linear projections are always beneficial for the conformational search or whether there are cases when usage of a projection might diminish the performance of a sampling-based algorithm.

Chapter 3

Background

In the current chapter we define two types of problems that conformational search addresses, introduce our framework for the conformational search, and show how it incorporates the apparatus of projection.

3.1 Conformational Search

In this thesis we considered two types of conformational search problems: the directed search and the undirected search. The characteristics of the produced conformational exploration that define whether the experiment succeeded or failed in terms of these problems are very different. Thus, the two presented problems required different approaches to handle them.

3.1.1 Directed Search

The first problem is a directed search. The main purpose of the directed search is to find a sequence of spatially-close low-energy protein conformations which constitute a feasible transition between given start and goal protein states as fast as possible. The analyzed in the thesis stochastic algorithm for the directed search is usually applied to the same input start and goal states several times to get a statistically representative

pathway or identify different pathways if more than one exists. In the problem of the directed search a sampling-based planner with some small probability (usually about 5%) tries to interpolate from a chosen low-energy state towards a goal state. This mechanism of biasing towards the goal conformation plays an extremely important role in how fast the sampling-based planner will discover the solution.

It worth noticing here, that due to the definition of the problem for the directed search, the sampling-based algorithm would favor the least amount of exploration of the conformational space. In the best case scenario, the algorithm explores only the direction towards the goal.

3.1.2 Undirected Search

The second problem considered is the undirected search: its purpose is to explore the protein conformational space starting from a given start protein state. In other words, the aim of the undirected search is to generate a variety of spatially-diverse conformations which would cover the subspace of a protein's low-energy conformations as fully as possible. In the contrast to the directed search problem, the undirected search does not use such a powerful tool to guide the exploration towards the feasible protein conformations as a goal bias. While we generally did not want to bias the undirected search *towards* any special direction, the ability to direct the undirected search *out* of the already sampled areas of the conformational space is vital for the success of the algorithm.

3.2 Structured Intuitive Move Selector

Our work is based on a framework for exploring the conformational space of proteins using a sampling-based motion planning approach called Structured Intuitive Move Selector (SIMS) [31]. The main purpose of SIMS is to investigate the space of low-energy conformations of a protein. SIMS operates in two modes: the directed or undirected search. For both modes, SIMS employs an advanced expansive sampling-based planner, and defines its main propagation procedures in terms of known protein *moves* (biophysically plausible perturbations of a protein's structure).

Protein Model

SIMS encodes a protein's conformation by the vector of its backbone dihedral angles. The angles between the bonds and their lengths are considered to be constant because they change insignificantly in comparison to the variation of the torsional angles. Side chains are optimized on-the-fly by the state-of-the-art Rosetta library [32, 33]. Such model has been shown to be a good enough approximation of a protein [34] and it drastically reduces the number of considered DoFs. Taking into account the planarity of the peptide bond, we restricted ω to be 180°. Thus, for each residue of the studied protein we kept only its (ϕ, ψ) values. This model induces 2N DoFs for a protein with N residues.

Fragments

Not all residues are equally important for a large-scale transition of a protein. Often only very few flexible parts of a protein are actively involved in its motion. SIMS is designed to allow for the prioritization of the most "active" parts of a protein: the algorithm gives more computational time to the exploration of these flexible regions. For this purpose, SIMS represents a protein as a set of flexible fragments, which can be defined by an expert (by specifying the the residues range in the input schema file) or automatically (from a protein's secondary structure). Each fragment consists of one or several subsets of a protein's residues. Depending on which parts of the molecule are known to be the most "active" in the studied motion, the fragments are assigned probabilities to be chosen during the sampling procedure. If the fragment is defined by an expert, then the expert should also assign the corresponding probability for this fragment to be sampled. When the fragment is created automatically from the secondary structure of the studied protein (using Rosetta library), the algorithm assigns weight to this fragment by the following rule: weight equals to 1, if the fragment represents a loop region; 0.2 if it represents a beta-sheet; or 0.1 if the fragment is an alpha-helix.

SIMS Algorithm

SIMS samples the state space and grows a tree of low-energy conformations, where each edge represents a possible transition from the parent state to the child state. To

increase the chances of sampling a low-energy conformation, SIMS uses an expansive approach: the algorithm grows its tree by randomly choosing a conformation which already belongs to the tree and trying to expand from it by slightly perturbing some of its DoFs.

The search algorithm takes start and goal states, the maximum allowed energy, the minimal distance (resolution step) as an input. There are several main steps of this algorithm (pseudo-code is presented in Algorithm 1 on page 19): 1) Use a projection to identify a possible parent state to expand from (line 4); 2) Slightly perturb the chosen state in a specific way (propagation step; line 5); 3) Compute the energy of the newly produced conformation (line 6); 4)) Accept the conformation and update the tree accordingly if the energy is below a user-defined threshold, otherwise the state is discarded (lines 6-9).

At each propagation step (pseudo-code is presented in Algorithm 2 on page 19), SIMS samples some fragment with a user-defined probability and slightly perturbs the residues of that fragment. In order to perturb conformations in a biologically feasible way, the framework involves the most common protein moves, such as loop motion, rigid body motion (fix one end of a loop and move the other end), energy minimization, and random perturbation. All mentioned moves (except energy minimization) are applied at the fragment level (i.e., the move affects only the residues of the chosen fragment). To implement these moves and to produce fast and accurate energy computations, SIMS uses the Rosetta library.

Algorithm 1 Search (startState, goalState, minDist, E_{max}, timeout)

```
1: addToTree(startState)
2: lastState ← startState
3: while distance(lastState, goalState) > minDist AND time < timeout do
4:     parentState ← SampleParent()
5:     currentState ← Propagate(parentState)
6:     if Energy(currentState) < E_max then
7:         addToTree(currentState)
8:         lastState ← currentState
9:     end if
10: end while
11: return Tree
```

Algorithm 2 Propagate (*state*)

```
1: fragment ← SampleFragment()
2: move ←SampleMove()
3: newState ← Apply(state, fragment, move)
4: return newState
```

3.3 Projection-based Expansion Heuristics

At each step, a sampling-based planner with some random heuristic chooses a promising parent state for the expansion of the tree towards the unexplored areas of the conformational space (Algorithm 1: line 4). This step is essential for the overall success of the algorithm: to a large degree it defines the direction where the conformational tree will expand to. To enhance the overall exploration of a protein's conformational space a lever is needed to softly bias the search procedure out of the well-sampled areas. A low-dimensional projection becomes such a lever for expansive planners. A planner uses a low-dimensional projection to keep track of the exploration progress by projecting each discovered feasible conformation on a low-dimensional subspace. The projection subspace is discretized into cells. The density of projected conformations in different cells estimates the density of the protein states in the corresponding areas of the conformational space. Based on this estimate the planner picks the state for the generation of a new conformation from the least populated projection cell (see Algorithm 3).

Algorithm 3 SampleParent()

1: $projectionCell \leftarrow$ PickCell()

2: $parentState \leftarrow$PickStateFromCell($projectionCell$)

3: **return** $parentState$

The particular way that the algorithm uses the projection to guide the exploration might vary depending on the implementation details of the employed plan-

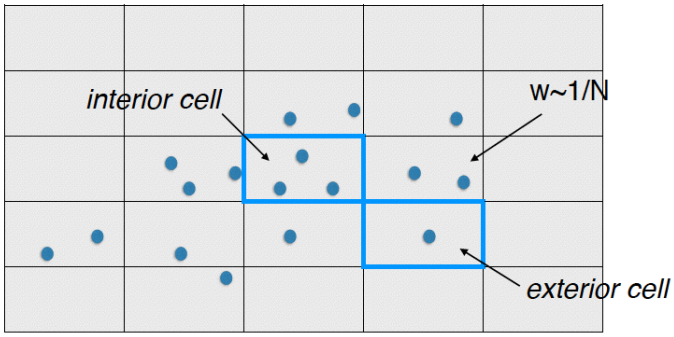


Figure 3.1 : 2D projection grid of the KPIECE planner distinguishes between interior and exterior cells. The probability of a cell to be picked is back-proportional to the number of conformations projected onto that cell.

ner. SIMS framework provides a connection to the Open Motion Planning Library (OMPL) [35], allowing an access to multiple OMPL planners for growing a graph of conformations. The level of abstraction implemented in SIMS makes it possible to use the majority of planners provided in OMPL. In the current work we focused on two particular planners, representatives of the expansive planner family: Expansive Space Tree (EST) [10] and Kinematic Planning by Interior-Exterior Cell Exploration (KPIECE) [11]. Both of these planners in the OMPL implementation rely on a linear projection to keep track of their explorational progress. In this section we will describe the main distinctions between these planners that cause some deviations in the performance of different kinds of projections for different kinds of problems (directed or undirected search).

EST is one of the forerunners of the expansive planners and has the most basic heuristic for choosing the next state for an expansion. At each step of the algorithm, EST picks one projection cell with the probability inversely proportional to the num-

ber of conformations projected onto that cell. From the chosen cell EST then picks the conformation at random with a uniform distribution.

KPIECE has a more comprehensive heuristic for choosing the next parent state. Firstly, KPIECE distinguishes between so-called "interior" and "exterior" cells. An exterior cell does not have a directly-attached, non-empty neighboring cell on at least one of its sides (Figure 3.1). The algorithm gives a large preference to the exterior cells for its further expansion. Secondly, once KPIECE has picked a cell, it then chooses the conformation from that cell randomly with a half normal distribution favoring the most recently added conformations.

Chapter 4

Construction of Projections

In the current chapter we describe the procedure of constructing a projection and define the main features of the successful projections. First, we design expert projections that take into account biological insights about a given protein. More specifically, we use the main "active" residues of the protein. We then present the algorithm for generation of a successful projection automatically based on the identified concepts of good projections. Finally, we introduce the benchmark protein systems with their expert-built projections.

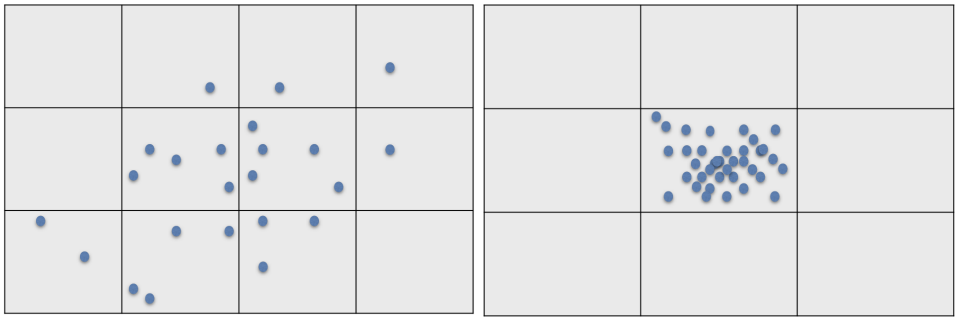
4.1 General Projection Methodology

The procedure of projecting a conformation was performed by multiplying its initial vector by a projection matrix. The conformations of a protein with N residues have $2N$ variables: (ϕ_i, ψ_i) for each residue i. Before applying the projection, we first transformed the conformation vector into a vector of sines and cosines: $(\phi_i, \psi_i) \rightarrow (\sin(\phi_i), \cos(\phi_i), \sin(\psi_i), \cos(\psi_i))$. This step transferred angular data to Euclidean space and allowed reasoning about projected conformation points in terms of Euclidean distances. Finally, the produced $4N$-dimensional vector was projected into a k-dimensional subspace by multiplying it by the projection matrix of size $k \times 4N$.

The projection space was discretized into a grid of equal-sized cells. This way, a projected point fell into some cell of the k -dimensional grid. The planner kept track of the number of conformations projected on each cell of this grid. The algorithm prioritized cells based on the density of coverage in different parts of the projection grid. At each iteration, the planner chose the highest-priority cell and randomly picked a state from this cell.

4.2 "Good" and "Bad" Projections

The intuition behind the technique of approximating a high-dimensional space with a low-dimensional projection was inspired by the Johnson-Lindenstrauss theorem [36]. This theorem states that distances between points in the initial n-dimensional space can be estimated with $(1 + \epsilon)$ distortion by the distances between the corresponding points embedded into a $\log(n/\epsilon^2)$-dimensional subspace. However, in the case of protein modeling, the dimensionality of the employed projection is usually much less than $\log(n)$. In general, the computational cost of maintaining a projection as well as the required memory resources grow exponentially with the number of dimensions. Therefore, in most cases the projection has no more than 10 dimensions; most often it has just 2 or 3. Because of such a significant reduction not all low-dimensional projections are able to estimate the initial high-dimensional conformational space equally well. Figure 4.1 illustrates how the *same* conformational graph can be projected on two different projection spaces. Figure 4.1a shows a good projection that can distin-



(a) Good projection (b) Bad projection

Figure 4.1 : An example of good and bad projections. **Good projection**: better distinguishes diverse states; better estimates coverage of the conformational space. **Bad projection**: does not distinguish states; cannot estimate coverage of the conformational space.

guish between many diverse protein states and, therefore, better estimates coverage of the conformational space. Figure 4.1b demonstrates an example of a projection that is not able to distinguish between any protein states, and doesn't deliver any useful information to guide the exploration.

How can we increase the chances to construct a good projection rather than a bad one? In the particular case of protein modeling (as opposed to modeling a robotic articulated chain), there is an additional factor justifying the usage of a low-dimensional projection: although proteins represent extremely high-dimensional systems, only very few of their "effective" DoFs are involved in large-scale motions [15, 16]. Therefore, the projection constructed from the few vectors corresponding to these flexible parts of the protein could represent a good approximation of the initial high-dimensional conformational space.

On the way towards generating an efficient projection automatically, we first have developed and tested a methodology to manually define the rows of a projection matrix based on simple biological intuitions about the studied protein. In the next chapter we demonstrate that a projection with this design represents a good approximation of the initial high-dimensional system.

To construct an expert projection, we first identified the main flexible regions of the considered protein. Second, we tried to predict how correlated these regions are. If some parts of a protein move mostly in a correlated way, we encoded them together into one of the projection's dimensions (instead of putting them into different dimensions of the projection matrix). Getting such biological insights about the studied protein involves the use of information available in the literature, analysis of available datasets of conformations, as well as visual inspection of the protein's secondary structure. After identifying which regions should be present in the projection matrix and how they should be coupled, we were ready to build the matrix. In each row we assigned non-zero values only to the variables encoding regions that should be coupled. The rows were then normalized. Flexible regions presented in each row did not intersect (i.e. each residue might occur only in one row of the projection matrix). This construction process ensured the orthonormality of the produced matrix, which is important to preserving relative distances in the projection space.

By construction, an expert projection differs from a random projection in the way that it employs only some predefined groups of residues that do not intersect,

which results in a sparse projection matrix. A random projection, on the other hand, uses linear combinations of *all* residues in each dimension. We believe that such "spreading" of the same residues along multiple projection dimensions is one of the main factors that causes the difference in the behavior of the random and specifically constructed projections.

To ensure a comprehensive analysis of the generated expert projections, we also built "misguided" projections. A misguided projection has the same nature as an expert one: in each row it encodes only some groups of residues. In contrast to an expert projection, the residues of a misguided projection are chosen in some protein's parts that are anticipated to be mostly rigid. With such construction, the subspace of a misguided projection should be mostly orthogonal to the subspace of the protein motion. As a result, a misguided projection cannot approximate the conformation space well and is not expected to enhance the exploration process.

4.3 Automatic Construction of Projections

In the next chapter we demonstrate that the expert projections defined in the previous section can significantly enhance conformational exploration. Because of that, we used the concepts of constructing the expert projections as the basis for generating an efficient projection automatically. The algorithm needs to 1) identify and organize the main flexible regions, and 2) distribute such regions between dimensions without overlapping. To identify the flexible regions within the studied protein, we used a

mechanism incorporated in SIMS that divides a protein into active fragments (see Section 3.2). These fragments might be defined by an expert and/or automatically based on the secondary structure of the protein. Each fragment has a weight assigned to it. This weight represents the probability of the fragment to be chosen for generation of a new conformation (i.e., new conformation would differ from its parent conformation by only values of the residues which belong to the sampled fragment). Aside from its weight, a fragment's importance also depends on its length: the longer the flexible region the greater the chances that it moves actively. Thus each fragment of the studied system has a tuple (weight, length) associated with it. We sorted the fragments by these tuples using the lexicographical order. Similarly to expert projections, we wanted auto projections to use only the most important fragments (the ones that are the most probable to be involved in function-related protein motions).

After the fragments are sorted, we used the Algorithm 4 on page 30 to assign them to dimensions of a projection matrix. The Algorithm 4 takes the ordered queue of the fragments and number of dimensions as an input, and returns the generated projection matrix. First, the Algorithm 4 initializes the projection matrix (lines 1-5). The matrix has a number of rows equal to the desired dimensionality of the projection and a number of columns equal to the initial dimensionality of the system (4xN, where N is a number of residues, see Section 4.1). The algorithm assigns the fragments with the highest priority first. To preserve relations between fragments' priorities, the algorithm assigns more fragments with lesser weight to one dimension.

The more residues share the same dimension - the less weight each one of them has on the guiding process. For this purpose, the algorithm keeps adding the fragments to the current dimension until its length exceeds the length of the previous dimension or reaches the defined limit. If the algorithm runs out of given fragments it fills the rest of the dimensions with randomly generated short continuous fragments. The important feature of the presented algorithm is that the fragments constituting different dimensions do not overlap with each other, i.e., each residue can have non-zero value in only one of the generated dimensions. To ensure this feature, every time after adding a new fragment the algorithm checks whether it introduces duplicates, and removes duplicated residues from the last added fragment (line 16).

All the heuristics employed in the described algorithm for generation successful projections automatically were chosen to mimic the construction of expert projections, which were shown to benefit the conformational search further in this thesis.

Algorithm 4 BuildMatrix (fragments, dimNum)

```
1: for (i = 0; i < dimNum; i++) do
2:     for (j = 0; j < 4*residueNum; j++)  do
3:         matrix [i, j] = 0
4:     end for
5: end for
6: maxRowLen = 4*residueNum / dimNum;
7: prevRowLen = 0, currRowLen = 0
8: currentDim = 0
9: while (currentDim < dimNum) do
10:    while (currRowLen<=prevRowLen AND currRowLen<=maxRowLen) do
11:        if (!fragments.empty()) then
12:            newFrag ← fragments.pop()
13:        else
14:            newFrag ← generateRandomFrag()
15:        end if
16:        addNewFragment(newFrag, currentDim, matrix)
17:        checkForDuplicates(matrix)
18:        currentRowLen = rowLen(currentDim, matrix)
19:    end while
20:    currentDim ++
21: end while
22: return matrix
```

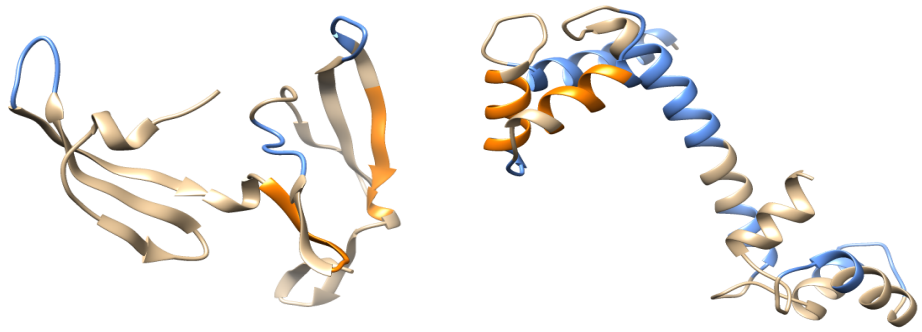
4.4 Studied Proteins and Associated Projections

For our experiments, we chose four well-studied protein systems with two known stable states: Cyanovirin-N, Calmodulin, Adelaide Kinase, and Ribose-binding protein. For each protein, we carried out a series of conformational searches to find possible low-energy transitions between these stable states (see Section 5.1). This experiment allowed us to evaluate the influence of the various projections on the success rate of the planner and on its average runtime. To assess their influence on the search-space coverage achieved by the planner, we also performed a series of conformational searches involving a single stable state (see Section 5.2).

Cyanovirin-N

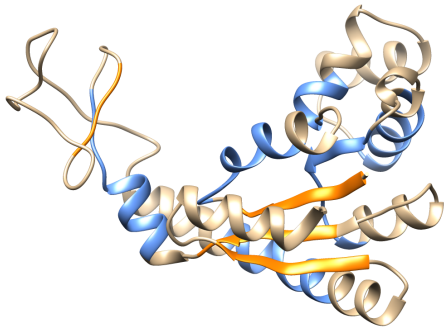
Cyanovirin-N (CVN) [37] is a bacterial protein with 101 residues, which corresponds to 202 DoFs in our framework. It demonstrates an antiviral activity towards several viruses including the human immunodeficiency virus (HIV). CVN is known to exist in two stable states, which can be found together in solution. To switch between these states, CVN goes through a domain swapping process, which involves a large-scale motion (the RMSD distance between the start state, PDB 2EZM, and the goal state, PDB 1L5E, is 17Å) via the correlated activity of three separate loop regions: residues 24-28, residues 50-55, and residues 75-80.

Based on this knowledge, we generated a three-dimensional expert projection matrix. Each row of this matrix encoded one of the mentioned loop regions by setting

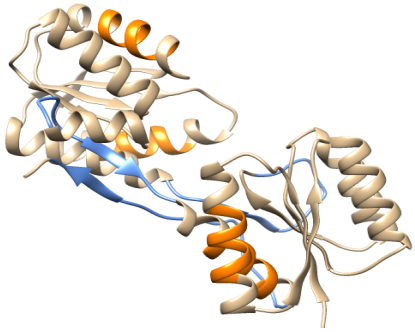


(a) Cyanovirin-N (CVN)

(b) Calmodulin (CaM)



(c) Adelaide Kinase (AdK)



(d) Ribose-Binding Protein (RBP)

Figure 4.2 : The four proteins involved in our experiments. Blue and orange areas indicate the residues involved in the expert and misguided projections, respectively.

only the elements corresponding to this loop's residues to non-zero values (represented by the blue regions in Figure 4.2a). The misguided projection also had three dimensions: the first two dimensions encoded residues 40-45 and residues 83-88, respectively (represented by the orange regions in Figure 4.2a), and the last dimension encompassed all the other residues. Residues 40-45 and 83-88 corresponded to the middle parts of beta-strands, which are likely to be inactive during the transition because of hydrogen bonds.

Calmodulin

Calcium-loaded Calmodulin (CaM) [38] is a middle-sized protein consisting of 144 residues (encoded by 288 DoFs in our framework). CaM is a calcium-binding protein involved in interactions between calcium ions and various target proteins. CaM exists in an open state (PDB 1CLL), and a closed state (PDB 1PRW) that are far apart from each other: the distance is about 16Å. The transition is known to happen mainly through unfolding of the middle part of the central helix.

Based on this information, we constructed a two-dimensional expert projection matrix. The first dimension contained the active residues of the central hinge (residues 67-80) [38]. The second dimension encoded regions of remaining active loops and some alpha helices involved in the transition (residues 5-20, 35-41, 52-57, 87-93, 107-116, 126-129). The misguided projection was generated from the residues of the alpha helices that were not involved in the main motion: both rows, the first and the second, contained residues 30-35, 47-52 (but with different signs in half of the values to ensure orthonormality of the matrix) (Figure 4.2b).

Adelaide Kinase

Adelaide Kinase (AdK) is a three-domain protein consisting of 214 residues, which corresponds to 428 DoFs. AdK catalyzes the transfer of a phosphoryl group from ATP to AMP, which is an important part of cellular energy homeostasis. The open (PDB 4AKE) and the closed (PDB 1AKE) states of AdK are about 7Å apart. The

exhaustive investigation of the forces causing the transitions between these states is presented in [39]. The work reveals the regions of high strain energy in AdK corresponding to residues 60-70 and 120-125, and to a lesser extent regions of residues 10-20, 30-35, 80-90 and 170-180. The authors show that the areas of the high strain drive the protein's opening and closing transitions by the process of local unfolding.

Incorporating this information into our expert projection, we created a three-dimensional expert projection matrix. The first two dimensions contained two fragments of the highest strain energy (first dimension: residues 60-70, second dimension: residues 115-125), and the third dimension aggregated all the secondary active regions (residues 10-20, 30-35, 80-90 and 170-180). The misguided projection was generated from the residues corresponding to the three beta-sheets buried in the stable core of the molecule (first dimension: residues 2-8, second dimension: residues 106-111, and third dimension: residues 131-134 and 192-197. See Figure 4.2c).

Ribose-binding protein

Our last system, Ribose-binding protein (RBP) [40, 41], is the largest presented in this work: it has 271 residues, which induces 542 DoFs in our model representation. RBP consists of two domains connected by three loop regions which form a hinge. The open conformation (PDB 2DRI) and the closed conformation (PDB 1URP) of this protein are only 4Å apart, but the transition between them requires a correlated motion of the three loop regions in the main hinge.

For this system we created a two-dimensional expert projection encoding the three loop regions of the hinge connecting the two domains. The first row contained two loop regions (residues 91-104 and 226-237). The second row corresponded to the third loop region (residues 253-269). Such choice of a placement of the flexible parts in the projection was made because the third loop region belonged to the very tail of the protein, and thus had more freedom for motions, whereas the first two loops were more constrained to move in a correlated way. The misguided projection was constructed from the residues of several alpha helices as follows: the first row contained residues 19-26 and 241-248; the second row contained residues 140-147 and 168-175 (Figure 4.2d).

Chapter 5

Assessment of Expert Projections

In this chapter we investigate whether we can improve the process of conformational sampling by employing expert projections (defined in the previous chapter) rather than using random ones. We compared the average performance of an expansive planner induced by an expert projection, as opposed to a random or a misguided projections. To ensure that the obtained results are not specific to a particular planner, we conducted the same experiments using two different expansive planners, EST and KPIECE.

5.1 Expert Projections for Directed Search

In the current section we describe the results obtained for a directed search problem (see Section 3.1.1). For each protein (with the exception of AdK), each type of projection, and two types of planners, KPIECE and EST, we performed 20 runs of a conformational search aimed at finding a feasible transition between a given pair of start and goal conformations. AdK system represents a significantly harder problem than that of the other considered proteins. For this reason, we have tested AdK system only with the more advanced KPIECE planner. Each experiment was held on a single thread of quad core 2.4 GHz Intel Xeon (Nahalem) CPUs with a 24-hour

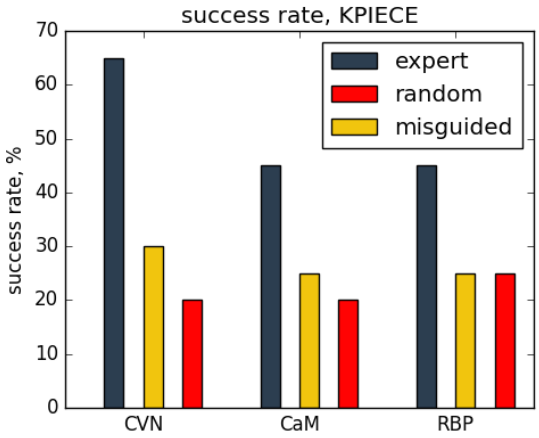


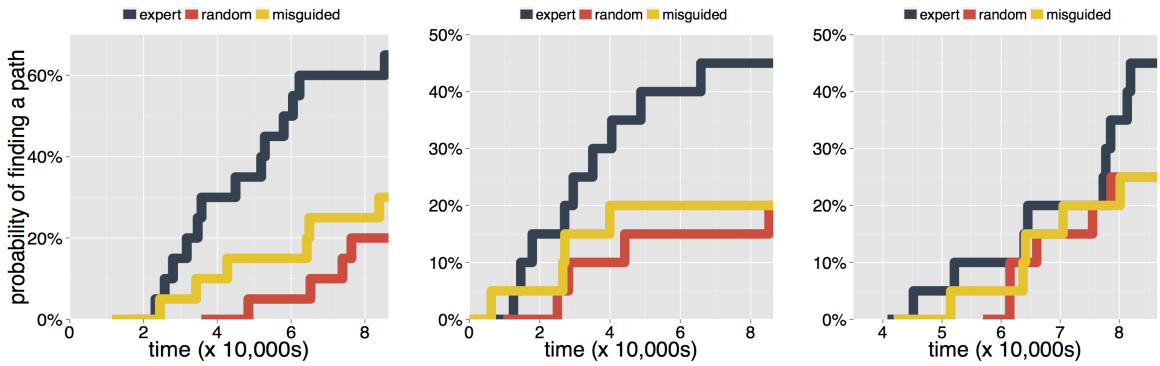
Figure 5.1 : Success rates associated with the three projection types: percentage of runs (among 20) by the KPIECE planner that successfully found a feasible transition between start and goal states using a particular type of projection (expert, random or misguided) for CVN, CaM, and RBP within a 24-hour time limit.

time limit. For both kinds of planners we obtained the consistent results of the expert projections producing more efficient exploration towards a goal state.

5.1.1 KPIECE Planner Results

For the CVN, CaM, and RBP proteins, we compared the success rates of the runs involving the expert, random, or misguided projections, respectively (see Figure 5.1). We define the success rate of a projection as the percentage of runs (among 20) that successfully find a feasible transition using that projection within a 24-hour time limit.

For CVN, the planner with the expert projection was 2.8 times more likely to find a solution than the planner with a random projection is, and 1.4 times more likely than the planner with the misguided projection is. Similar results were obtained for



(a) CVN (b) CaM (c) RBP

Figure 5.2 : Probability of finding a solution path by the KPIECE planner as a function of time for each of the three projection types: CVN, CaM, and RBP. Grey color is associated with an expert projection; red color - with a random projection; and yellow color - with a misguided projection.

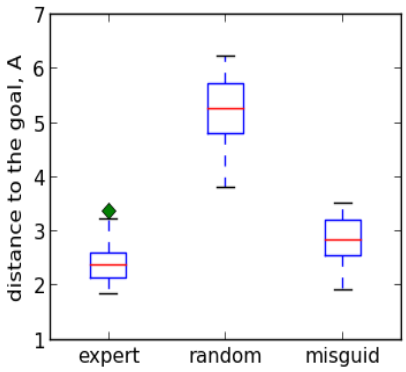


Figure 5.3 : Minimal distance to the goal achieved in runs (among 20 runs) by the KPIECE planner with three projection types (expert, random, and misguided) for AdK within a 24-hour time limit.

CaM: the expert projection was successful 2.3 times more often than the random projection was and 1.8 times more often than the misguided projection was. For RBP, the expert projection demonstrated 1.5 times more successful runs than those by the random and misguided projections.

Therefore, despite the differences between the three conformational search problems, the expert projections demonstrate a consistent improvement over the random and misguided projections, in terms of success rate. Supported by Figure 5.2 we also can state that the expert projections accelerate the sampling-based algorithm more than the other considered types of projections do. Figure 5.2 shows the probabilities of finding a solution path as a function of time for each of the three projection types. The success probabilities corresponding to the final time step, 86400 seconds (24 hours), in Figure 5.2 are the values of the overall success rates presented in Figure 5.1. Figure 5.2 demonstrates the fact that for each particular time step the expert projections have a higher probability of discovering the solution for each of the studied proteins.

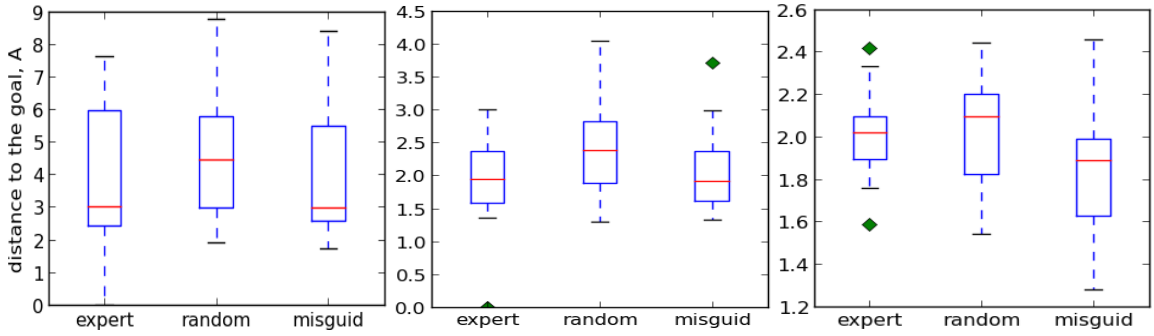
Finding a transition pathway for AdK protein is a harder problem because AdK is larger than the CVN and CaM proteins, the distance between its start and goal states is significantly greater than the distance in RBP problem, and the character of AdK motion has more complex non-linear nature. For these reasons, KPIECE runs have not been able to identify the exact solutions for AdK. However, we can evaluate how close to the goal the search has reached (Figure 5.3). The expert and misguided projections have been able to bring the conformational search much closer to the goal state than the random projection. The initial distance between AdK start and goal states is about 7Å. The expert and misguided projections have discovered conformations that are just 2.5 and 2.8Å away from the goal correspondingly. The

runs with the random projection, on the contrary, demonstrate very little progress toward the goal: they usually stop no closer than 5.5Å from the goal.

Figure 5.2 and Figure 5.3 illustrate that the expert projection consistently has a higher probability of finding a solution or getting closer to the goal state in a given time. As computational time is a very limited resource, especially for modeling large proteins, the usage of the expert projection can significantly benefit the simulations.

5.1.2 EST Planner Results

EST has a much simpler heuristic (see Section 3.3) than the KPIECE planner does. Because of this, the EST planner requires more computational time to guide the exploration from the start to the given goal state. In the conducted experiments for CVN, CaM, and RBP only very few EST runs were able to identify solution pathways. Because of this, we do not provide the success rate statistics for the EST planner. Instead, to assess different kinds of projections (expert, misguided, and random) with the EST planner we compared the progress towards the goal state achieved in 24 hours of computations by the runs with different projections for CVN, CaM, and RBP. Figure 5.4 shows that the runs with the expert projections are getting closer to the goal state in a 24-hour period than the random projections do. For CVN and CaM, only runs with the expert projections were able to find solutions (an expert projection has produced 1 solution out of 20 runs for CVN and 2 solutions out of 20 runs for CaM). Surprisingly, the performance of the misguided projections is



(a) CVN (b) CaM (c) RBP

Figure 5.4 : Minimal distance to the goal achieved in runs (among 20 runs) by the EST planner with the three projection types (expert, misguided and random) for the CVN, CaM, and RBP proteins within a 24-hour time limit.

very close to the performance of the expert projections. One of the possible reasons of this phenomena is the fact that even though the misguided projections possibly incorporate the most rigid parts of a protein, they still have the same nature as the expert projections do (their matrix is sparse; the segments of residues presented in each dimension do not intersect with each other) as opposed to that of the random projection.

5.2 Expert Projections for Undirected Search

In this section we compare the quality of exploration coverage produced by the planners using different types of projections. To explore the conformational space as vastly as possible, we performed an undirected search (see Section 3.1.2). The purpose of this search is an extensive exploration of the conformational space of a protein starting from a given state (in other words, no goal state is involved). This way, we

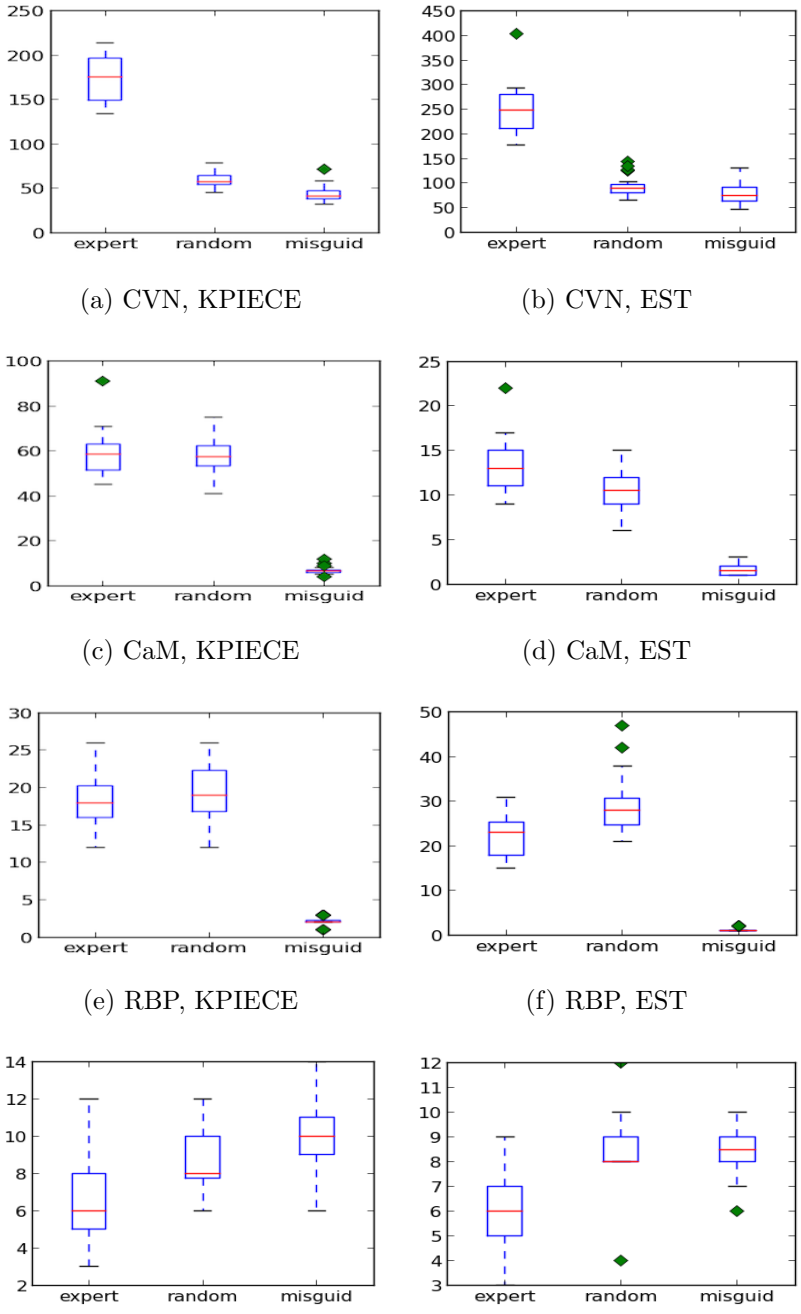
are not exploring only the protein flexibility inherent to a single transition, but the overall flexibility of the protein. In this experiment, we performed 20 runs of this conformational search for each protein, each type of projection, and two types of planners (EST and KPIECE). Each experiment was held on a single thread of quad core 2.4 GHz Intel Xeon (Nahalem) CPUs with a 24-hour time limit.

5.2.1 Coverage of the Projected Space

One way to quantify the influence of the projection on the process of the undirected conformational search is to quantitatively assess the amount of explored projection space. A good alignment of guiding projections with a protein's flexibility results in increased volume of the explored projection space. The large volume of the explored projection space indicates that the projection is able to distinguish a lot of different conformations and assign them to different cells. We are interested in increasing the volume of the explored *projection* space, because this potentially translates into enlarging the volume of the explored *conformational* space, as well as rises the overall projection's influence on the search algorithm.

It is important to note that a small number of explored projection cells does not necessarily indicate a bad exploration of the conformational space. On the other hand, a large number of non-empty cells is an indicator of good conformational space coverage.

Figure 5.5 illustrates the average number of projection cells explored during the



(a) CVN, KPIECE (b) CVN, EST

(c) CaM, KPIECE (d) CaM, EST

(e) RBP, KPIECE (f) RBP, EST

(g) AdK, KPIECE (h) AdK, EST

Figure 5.5 : Average number of projection cells explored by KPIECE and EST during the conformational search with each type of projection within 24 hours for CVN, CaM, RBP, and AdK.

conformational search with each type of projection for KPIECE and EST respectively. The results produced by two planners are very consistent with each other. In both cases for CVN system, the searches with the expert projection discovered three times more projection cells than the searches with the random or misguided projections. For CaM and RBP, runs with the expert projections explored a volume of projection space similar to what runs with the random projections did, but much greater than what runs with the misguided projections did. The larger and more constrained AdK system yielded very few projection cells explored for all three considered types of projections.

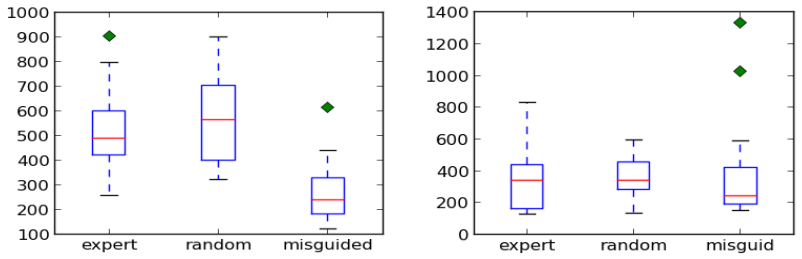
Even though the expert projections described in Section 4.4 do not incorporate all flexible parts of the studied proteins (they employ only the flexible parts that are anticipated to be involved in the transition between the start and goal states), the coverage of the projection space they produce is at least as good as the one produced by a random projection and, sometimes, better (Figure 5.5). The poor coverage produced by the misguided projections was expected. The misguided projections were designed specifically to focus mostly on rigid parts of a given protein. As a result, the generated low-energy conformations of the studied protein are likely to be distributed along the directions that are mostly orthogonal to its projection space. In this set of experiments, a random projection demonstrates relatively good performance, especially for flexible proteins. In the case of flexible proteins, there is a higher chance to randomly generate a projection that induces good exploration of the projection space

because almost any combination of residues could be involved in *some* motion (not necessarily function-related).

5.2.2 Coverage of the Conformational Space

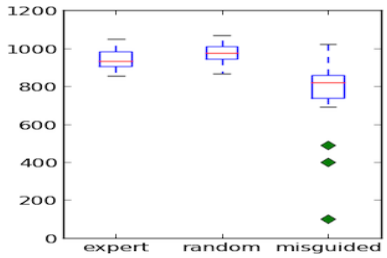
Coverage of the projected space represents a major indicator of how well the projection is aligned with the overall protein flexibility, but it does not necessarily produce an accurate picture of the *conformational* space coverage. To estimate the volume of the explored conformational space we used the mechanism of covering the produced protein conformations with 2N-dimensional balls (2N is the dimensionality of the protein system with N residues) of a fixed radius r. We greedily computed a ball coverage by randomly picking a conformation from the generated tree of conformations, which becomes a center of a new ball. Then we computed the distance to the k-nearest neighbors of the chosen conformation and removed from the considered set those of them that were within a distance r. We repeated the described process until we ran out of conformations. Finally, we counted how many balls of radius r were produced by the algorithm. Even though, the described algorithm computes a *random* coverage, rather than *minimal*, its standard deviation is usually less than 1%. Therefore, this metric gives us a good approximation of the explored conformational space volume.

As it was mentioned in the previous section, a random projection is expected to produce more expansive coverage than an expert projection, because the expert pro-

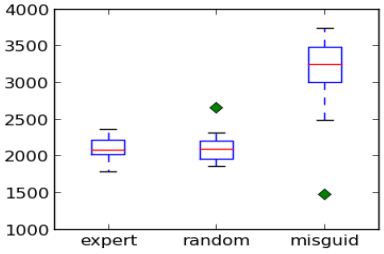


(a) CVN, KPIECE

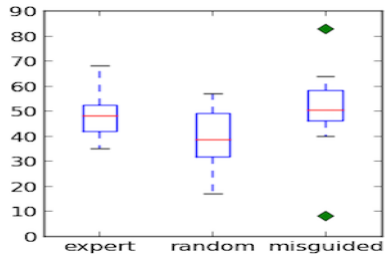
(b) CVN, EST



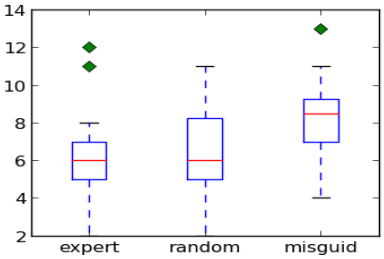
(c) CaM, KPIECE



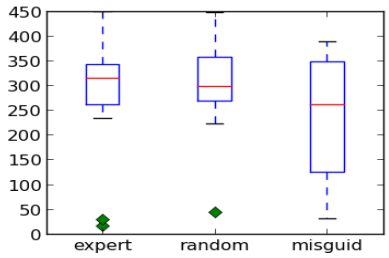
(d) CaM, EST



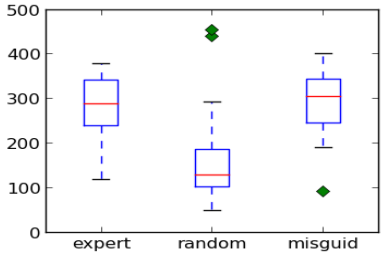
(e) RBP, KPIECE



(f) RBP, EST



(g) AdK, KPIECE



(h) AdK, EST

Figure 5.6 : The number of balls with a radius r=1Å produced by the expert, random, and misguided projections in 24 hours, for CVN, CaM, RBP, and AdK using KPIECE and EST planners.

jection was designed to keep track only of a protein's parts that are anticipated to be involved in the *function-related* motions, rather than *all* possible motions. This way the expert projection is supposed to guide the exploration towards the anticipated goal direction, and ignore the other directions. However, Figure 5.6 demonstrates that the expert projection produces similar or better coverage than the random projection does. The misguided projections usually produce equal or less conformational exploration than the expert and random projections do, except for the case of CaM problem, where the EST planner with the misguided projection has generated 1.5 times better explorational coverage than the other projections. One of the possible explanations for such a result is the fact that CaM is very flexible protein. Even though the misguided projection does not incorporate the flexible parts of CaM that are anticipated to be involved in the start-goal transition, the residues that it does employ might be significant for several other possible directions.

5.3 Summary of the Results

In this chapter we have investigated how the expert-defined projections that employ available knowledge of a protein's behavior affect the sampling-based exploration process. We have shown that the projections designed to incorporate the information related to a single start-goal transition are the most useful for the directed search problem, where they significantly increase the probability of the algorithm to find the studied transition in the given time. In terms of the undirected search problem, when

the algorithm needs to explore all possible feasible motion directions of a protein, all three considered types of projections demonstrate very similar results. Therefore, the usage of a random projection might be good enough for the purpose of undirected exploration. Alternatively, the concepts of expert projection might be modified to suit the purpose of the undirected search: the projection should incorporate *all* flexible protein regions rather than only those that are active during one specific transition.

Chapter 6

Assessment of Automatically Generated Projections

In the previous chapter we have shown that employment of special expert projections can significantly improve the performance of the expansive sampling-based planners for the directed search problem. However, the construction of an expert projection is a tedious process requiring a lot of preprocessing work to obtain additional knowledge about the studied protein and generation of a projection matrix manually. To eliminate the preprocessing steps but still be able to take advantage of good projections, we developed an algorithm for construction of reasonable projections automatically (see Section 4.3). In the current chapter, we evaluate the performance of the projections constructed automatically by the proposed algorithm. Similar to the assessment of the expert projections in the previous chapter, we evaluate the performance of the auto projections on two benchmark problems: the directed (start-goal) and undirected conformational searches.

6.1 Auto Projections for Directed Search

In this section we describe the results obtained for a directed search problem. In addition to the runs with the expert, random, and misguided projections described

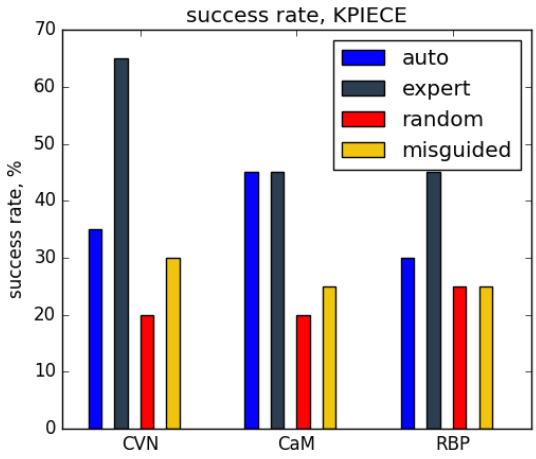
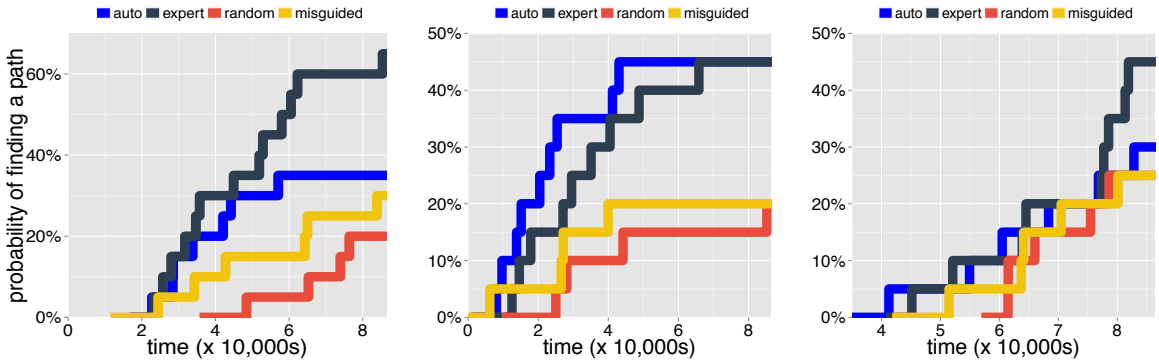


Figure 6.1 : Success rates associated with the four projection types: percentage of runs (among 20) by the KPIECE planner that successfully found a feasible transition between start and goal states using a particular type of projection (auto, expert, random or misguided) for CVN, CaM, and RBP within a 24-hour time limit.

in the previous chapter, we performed 20 runs of the directed search with EST and KPIECE planners using the projections generated automatically. Each experiment was held on a single thread of quad core 2.4 GHz Intel Xeon (Nahalem) CPUs with a 24-hour time limit.

Figure 6.1 and Figure 6.2 show the success rates induced by the four considered projection types with KPIECE planner for CVN, CaM, and RBP systems. Even though the auto projections usually do not accelerate the conformational search as well as the expert projections do, they still noticeably increase the probabilities of finding a solution path compared to the random and misguided projections. For CaM, the auto projection even outperforms the expert projection (Figure 6.2b). demonstrated by The auto projection with the KPIECE planner for AdK system demonstrates very close performance to the one produced by the expert projection (Figure 6.3). Simi-



(a) CVN (b) CaM (c) RBP

Figure 6.2 : Probability of finding a solution path by KPIECE planner as a function of time for each of the four projection types, for CVN, CaM, and RBP. Blue color is associated with an auto-generated projection; grey color - with an expert projection; red color - with a random projection; and yellow color - with a misguided projection.

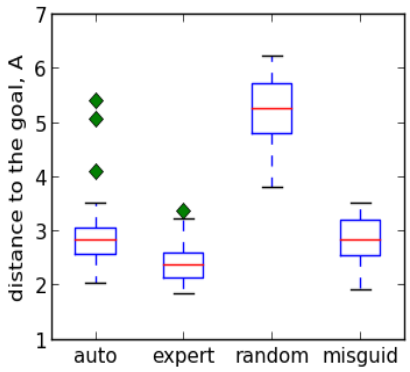
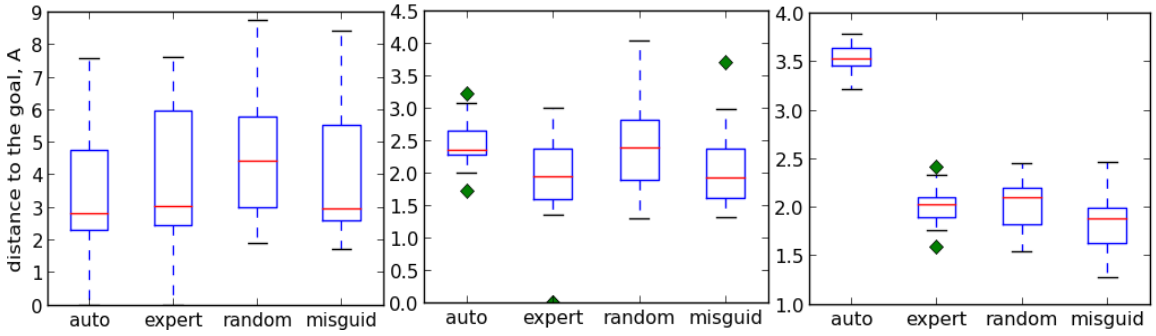


Figure 6.3 : Minimal distance to the goal achieved in runs (among 20 runs) by KPIECE planner with the four projection types (auto, expert, random, and misguided) for AdK within a 24-hour time limit.

lar to Chapter 5, we did not test AdK system with the EST planner as this system might be too complex for EST to generate meaningful results in 24 hours. For the other studied proteins, EST planner produces close to KPIECE results (Figure 6.4). Although the auto projection does not perform very well with EST planner in RBP



(a) CVN (b) CaM (c) RBP

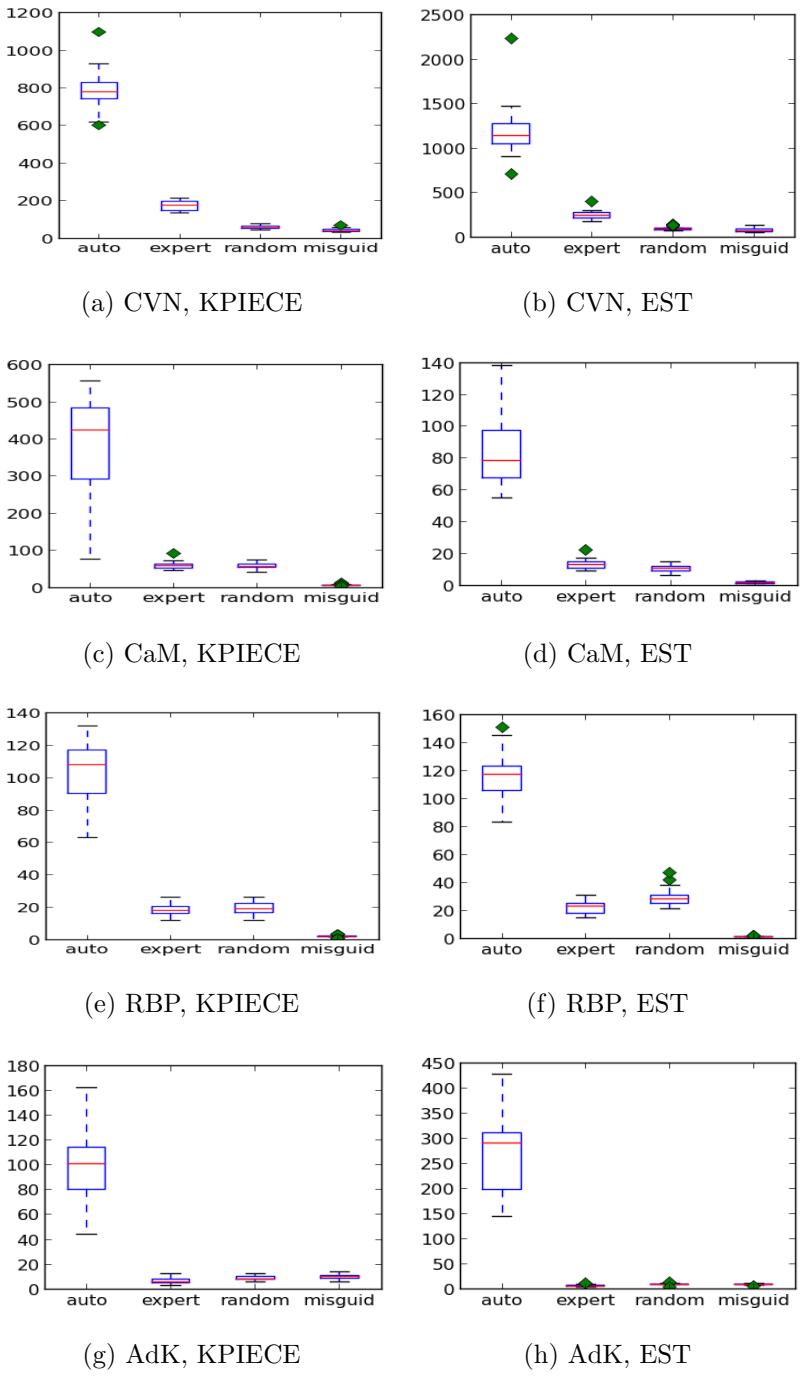
Figure 6.4 : Minimal distance to the goal achieved in runs (among 20 runs) by the EST planner with four projection types (auto, expert, misguided and random) for the CVN, CaM, and RBP proteins within a 24-hour time limit.

problem, it still enhances the search for all the other studied systems.

Overall, as expected, the auto projection does not always perform as well as the expert projection does, but in the most cases it presents a significantly better alternative to the random projection.

6.2 Auto Projections for Undirected Search

In this section we assess the properties of the exploration produced by the automatically generated projections in the undirected search benchmark. In this experiment, we performed 20 runs of this conformational search for each protein, each type of projection, and two types of planners (EST and KPIECE). Each experiment was held on a single thread of quad core 2.4 GHz Intel Xeon (Nahalem) CPUs with a 24-hour time limit.



(a) CVN, KPIECE (b) CVN, EST

(c) CaM, KPIECE (d) CaM, EST

(e) RBP, KPIECE (f) RBP, EST

(g) AdK, KPIECE (h) AdK, EST

Figure 6.5 : Average number of projection cells explored by KPIECE and EST during the conformational search with each type of projection within 24 hours for CVN, CaM, RBP, and AdK.

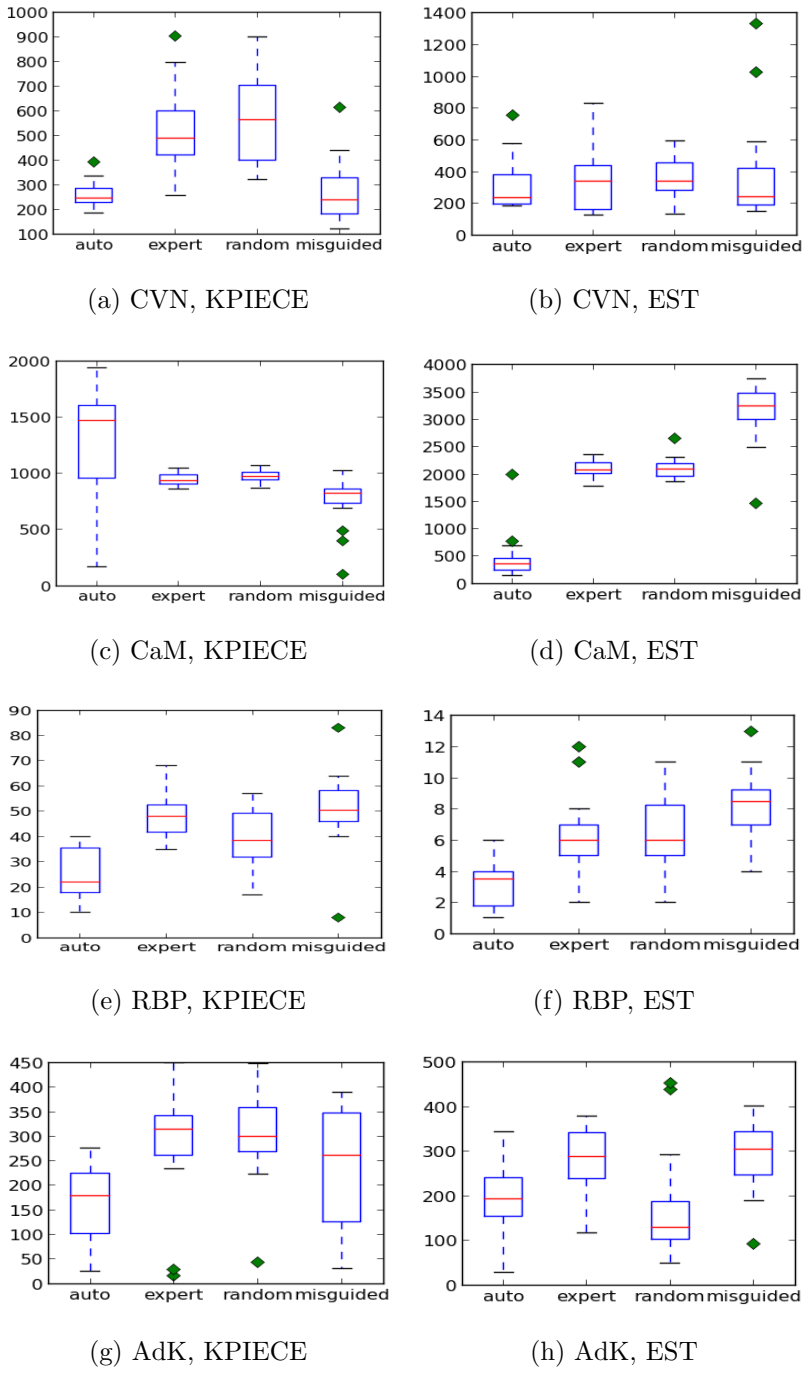
6.2.1 Coverage of the Projected Space

The auto projection demonstrates significant increase in the volume of the explored projection space consistently with all four studied systems and both considered planners (Figure 6.5). This result indicates that the algorithm defined in Section 4.3 manages to produce projections that are very well-aligned with the overall protein flexibility and are able to differentiate between many diverse protein states.

6.2.2 Coverage of the Conformational Space

To estimate the conformational space coverage produced by the auto projections, we employed the technique (introduced in Section 5.2.2) of covering the generated tree of the protein conformations with 2N-dimensional balls. The obtained results are presented in the Figure 6.6. The auto projection produces mixed results for this measure. Like the results of the directed search problem, the best performance of auto projection is achieved for the CaM protein with the KPIECE planner. The exploration of conformational space of CaM induced by the KPIECE planner with the auto projection exceeds the similar measurements generated by the other types of projections by 1.5 times. For the other considered proteins, the auto projections yield similar or less exploration than the other projections.

The employed algorithm for the automatic generation of a projection relies on the main concepts of the expert projection construction. More precisely, the produced auto projection incorporates only a protein's fragments with the highest weight (the



(a) CVN, KPIECE (b) CVN, EST

(c) CaM, KPIECE (d) CaM, EST

(e) RBP, KPIECE (f) RBP, EST

(g) AdK, KPIECE (h) AdK, EST

Figure 6.6 : The number of balls with a radius r=1Å produced by the auto, expert, random, and misguided projections in 24 hours, for CVN, CaM, RBP, and AdK using KPIECE and EST planners.

most flexible and long fragments as defined in Section 4.3), and not *all* possible flexible parts. The automatic procedure might miss some important flexible parts of a protein, especially if such parts appear in the regions of alpha-helices, which are considered to represent the most rigid regions of a protein's secondary structure, but nonetheless they are very often involved in a protein's critical motions as in case of CaM and AdK. Due to these limitations, the auto projection constructed in this work was not expected to produce a large coverage of the conformational space for all proteins.

6.3 Summary of the Results

The mechanism for constructing of a good projection automatically, as defined in the current work, aims to approximate the expert projection. The way we constructed the expert projections (incorporating only the residues that are anticipated to be important for a *single* start-goal transition) made them the most useful for the directed search problem (see Figure 5.2), but did not add much improvement for the undirected search problem (where it is beneficial to explore *all* possible low-energy transitions). Therefore, similar to the expert projection, the auto projection is the most useful in the directed search problem. Although the auto projection does not have an exact information about which parts of the considered protein are the important and thus is less accurate than the expert projection, it still outperforms the random and misguided projections in terms of the directed search problem for all four

studied proteins. In some cases, the auto projection is even more successful than the manually-produced expert projection (as in case of CaM).

Chapter 7

Discussion on the Applicability of Projections

In this chapter we investigate the question of whether the linear projections are always useful for the conformational search or if cases exist when the wrong projection could diminish the performance of the algorithm. We compare the performance and the quality of the exploration coverage produced by the runs with a bias of a projection (expert or random) and without any projection guidance (a single-cell projection). We also describe a hypothesis explaining the produced results. However, we are still working on generating a solid experimental base to support the given hypothesis.

7.1 Search Using No Projection

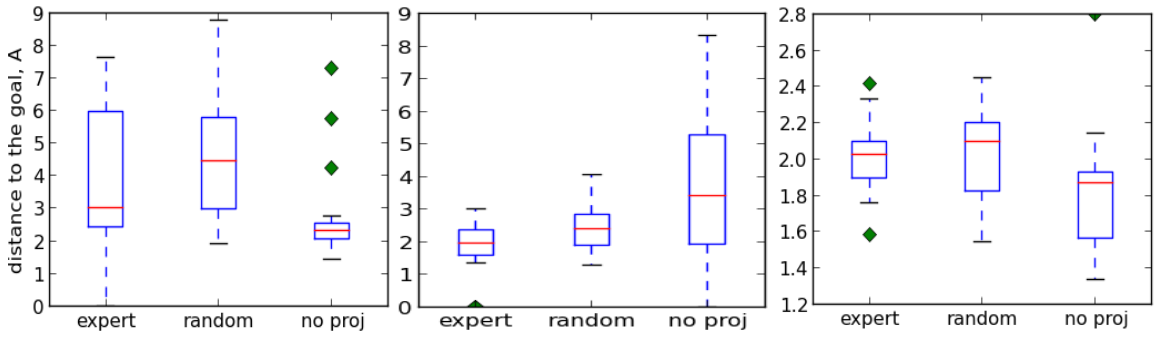
The surprisingly good performance of the misguided projections in the section 5.1 has raised a question. The misguided projections were developed specifically to reduce the possible effect of a projection on the overall algorithm performance. For this reason, the construction of the misguided projections had ensured that they will generated very few projection cells (see Figure 5.5). Employing the concept of the misguided projection in an extreme case, we generated a projection that would project the whole conformational graph onto a single cell in the projection space. This way we completely eliminated the effect of the projection on the exploration process: the

algorithm's success or failure depends only on its heuristic of choosing the conformation from the single cell. In the current chapter we examine EST and KPIECE single-cell heuristics (in the OMPL implementation) for guiding the exploration, and compare them to the projection-biased exploration (case when a projection generates multiple cells). EST chooses a conformation from a given cell uniformly at random; KPIECE picks a conformation from the cell using a half-normal distribution favoring the most recently-added conformations (see Section 3.3). In the next sections, we will demonstrate that this distinction plays an important role in the planners' performance and quality of exploration with or without projection guidance.

As our aim in this chapter is to compare a no-projection exploration to a possible projection-based exploration (not necessarily best or worse projection exploration), we do not consider the auto and misguided projections in the experiments for the current chapter.

7.2 Impact of Projections on Directed Search

In this section we present the results obtained for a directed search problem. For CVN, CaM, and RBP, we performed 20 runs of a conformational search aimed at finding a feasible transition between a given pair of start and goal conformations using expert, random, and single-cell projections with KPIECE and EST planners. For AdK problem, because of its complexity, we performed the similar experiments only with the KPIECE planner. Each experiment was held on a single thread of quad



(a) CVN (b) CaM (c) RBP

Figure 7.1 : Minimal distance to the goal achieved by the EST planner in runs (among 20 runs) with the different projection types (expert, random, and single-cell) for CVN, CaM, and RBP within a 24-hour time limit.

core 2.4 GHz Intel Xeon (Nahalem) CPUs with a 24-hour time limit.

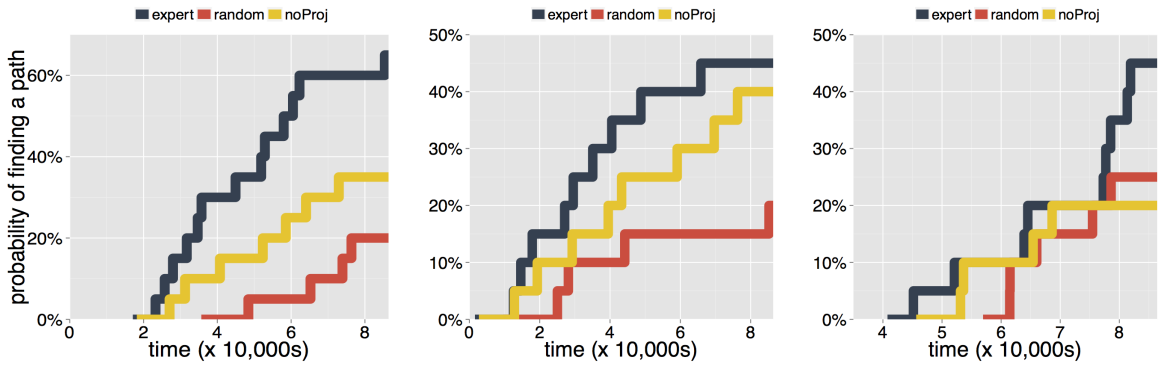
Figure 7.1 presents the performance comparison of the EST planner with multi-cell projections (expert, random) and single-cell projection for the CVN, CaM, and RBP proteins. CVN and RBP exhibit the similar results, whereas CaM demonstrates the mirroring to the other two proteins performance.

For the CVN protein, only the expert projection was able to identify a solution (1 solution out of 20 runs), but the overall multi-cell projection results exhibit a lot of inconsistency, whereas the runs with the single-cell projection much more consistently bring the conformational search close to the goal (Figure 7.1a). This result is surprising taking into account the fact that EST with the single-cell projection on each step chooses the conformation for the further expansion absolutely randomly with the uniform distribution, whereas any multi-cell (expert, random, or any other) projection was expected to create a bias towards a goal state and thus be more consistent.

For the CaM protein, we had an opposite results (Figure 7.1b): the single-cell projection yields much more variation in the results than a multi-cell (expert or random) projection does, but at the same time, it is more successful at finding solution pathways (single-cell projection has generated 5 solutions out of 20 runs, compared to just 1 solution for the expert projection).

In case of the RBP system, the single-cell projection produces the pathways that on average are closer to the goal state than the pathways induced by multi-cell projections. This result is consistent with the performance obtained for the CVN system.

One of the possible explanations for the difference in the behavior of CaM compared to CVN and RBP is the fact that CaM is a very flexible protein, whereas the other two proteins are rather rigid. In other words, CaM has various distinct low-energy directions, whereas CVN and RBP have very few (possibly only one) feasible directions. We believe that the mechanism of a projection interferes with the efficient growth of the conformational tree for the *highly constrained* systems (such as CVN and RBP), but might be useful for the exploration of the *flexible* systems (such as CaM). At each new step, a multi-cell projection tries to pick the different cell for further expansion, even if the previous expansion successfully produced a new conformation. Such an approach would slow down the exploration of systems with a single narrow passage (Figure 7.1a, 7.1c): instead of proceeding the identified promising direction, a projection forces the algorithm to abandon the current cell, and explore new possible directions. A projection makes the algorithm forget where the narrow



(a) CVN (b) CaM (c) RBP

Figure 7.2 : Probability of finding a solution path as a function of time for each of the three different projection types, for CVN, CaM, and RBP. Dark blue color is associated with an expert projection; red color - with a random projection; and yellow color - with a single-cell projection (no projection).

passage is and search for it again at each step. For the flexible proteins, such an approach will result in a very thorough exploration of the low-energy conformational landscape (a projection makes the algorithm grow the conformational tree in all possible directions at the same time), which might slow down the search for a particular trajectory, but will generate very consistent results (Figure 7.1b).

The additional projection heuristics (see Section 3.3) help the KPIECE planner to use the mechanism of projection more successfully. However, the obtained results (Figure 7.2 and 7.3) suggest that the projection needs to be chosen very carefully: a random projection often significantly diminishes the performance of the algorithm.

The described reasoning still represents a hypothesis, and we are actively working on the building the basis of support for it.

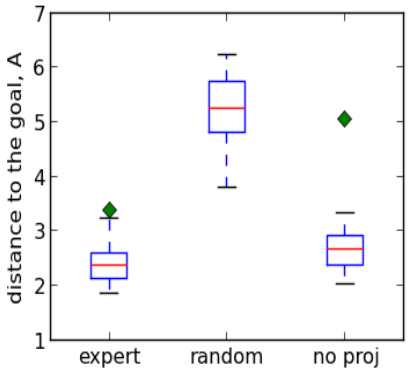
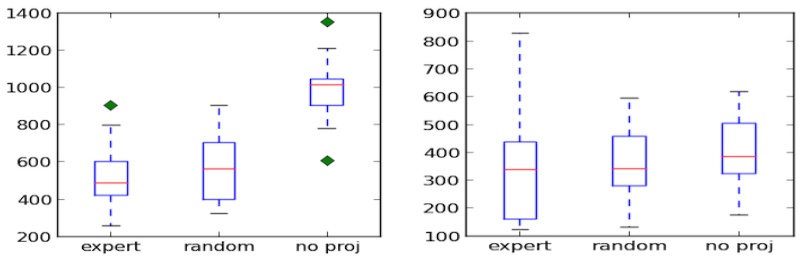


Figure 7.3 : Minimal distance to the goal achieved by the KPIECE planner in runs (among 20 runs) with the different projection types (expert, random, and single-cell) for AdK within a 24-hour time limit.

7.3 Impact of Projections on Undirected Search

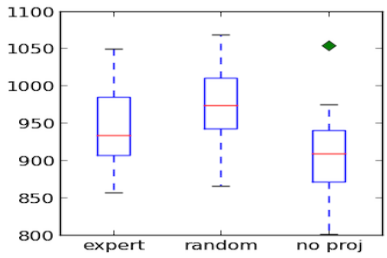
In this section we quantify the conformational exploration produced by multi-cell and single-cell projections during the process of the undirected search. In this set of experiments, we performed 20 runs of the undirected conformational search for four proteins (CVN, CaM, RBP, and AdK), for each type of projection (single-cell or multi-cell), and two types of planners (EST and KPIECE). Each experiment was held on a single thread of quad core 2.4 GHz Intel Xeon (Nahalem) CPUs with a 24-hour time limit.

To evaluate the volume of explored conformational space we compute coverage of the produced conformational graphs with 2N-dimensional balls (see Section 5.2.2). Figure 7.4 presents the obtained results. Comparing Figure 7.4a to Figure 7.4b, Figure 7.4c to Figure 7.4d, and Figure 7.4g to Figure 7.4h we can see that for the same protein system EST and KPIECE planners agree with each other on whether it

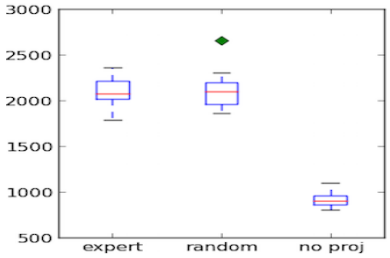


(a) CVN, KPIECE

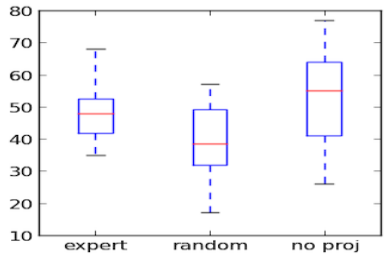
(b) CVN, EST



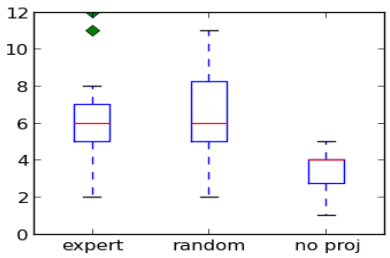
(c) CaM, KPIECE



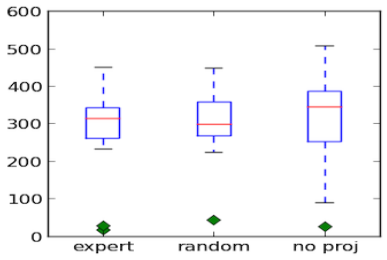
(d) CaM, EST



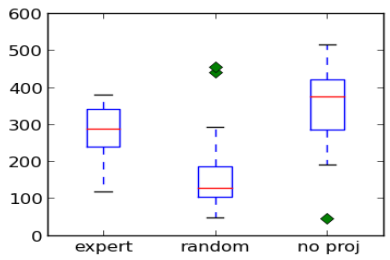
(e) RBP, KPIECE



(f) RBP, EST



(g) AdK, KPIECE



(h) AdK, EST

Figure 7.4 : The number of balls produced by expert and random projections in 24 hours with a tile's diameter d=1Å, for CVN, CaM, RBP, and AdK using the KPIECE and EST planners.

is worth using a multi-cell or a single-cell projection. However, due to the difference in the planners' heuristics one of the planners usually has a more pronounced distinction in results produced by these projections than the other.

For CVN, usage of a single-cell projection with the KPIECE planner increases the planner's productivity by 2.5 times compared to a multi-cell expert projection (Figure 7.4a). At the same time, if the EST planner is used to explore CVN conformational landscape, then employing both types of projections produces on average very close results (Figure 7.4b).

A similar exploration pattern is presented by the other constrained protein, AdK. Although, in case of KPIECE both types of projections generate a very similar exploration volume for AdK (Figure 7.4g), the exploration by the EST planner noticeably benefits from the single-cell projection (Figure 7.4h).

RBP also represents a highly-constrained protein, and the obtained for RBP results are consistent with CVN and AdK results.

CaM illustrates the example when the multi-cell projection drastically increases the quality of conformational exploration. Figure 7.4d shows how the projection mechanism of growing the exploration tree in all possible directions at once might greatly enhance the exploration of a flexible system (system with many feasible directions to explore).

7.4 Discussion of the Results

To explain the observed results we developed a hypothesis of how a multi-cell projection works and affects the conformational tree. The concept of a projection was developed specifically to keep track of already explored parts of the conformational space, and drive the exploration away from them. In practice, such mechanism forces a planner to grow the conformational tree in all possible directions at once. This happens because when the algorithm successfully generates a new child conformation from the known parent one, the child state is usually very close to its parent, and thus it most probably will get into the same projection cell as its parent. Therefore, the density of the current projection cell increases, which leads the algorithm to choose another cell in the next step, and explore another possible direction. Thus, we would argue that a projection produces an "exploration-first" search, rather then a "refinement-first" search. The exploration-first approach might be beneficial for the purpose of the undirected exploration for the flexible structures, when there exist a lot of possible directions and the algorithm is forced to explore them all (Figure 7.4d). However, for the constrained systems, when there is only one or very few feasible directions, the exploration-first approach that in each step forces the exploration to leave the encountered narrow passage might significantly harm the performance (Figure 7.4a, 7.4h). Similarly, for the directed search problem, when the algorithm is looking for a single feasible trajectory, the exploration-first search could delay the discovering of a solution. However, in the case of the directed search problem the algorithm also has

a strong guiding force of the goal-bias that partly compensates the losses caused by a projection. Applied together, a good projection and a goal-bias could even accelerate the process of finding a feasible start-goal trajectory (Figure 7.2).

The performed investigation of the applicability of the linear projections has revealed the surprising fact that projections do not always enhance the exploration. When deciding whether to use a projection at all, one needs to take into consideration which type of a problem (directed or undirected search) is being solved, whether enough information about the studied protein is available to construct a good projection, and whether the protein is rigid or flexible among other possible factors.

Chapter 8

Conclusions and Future Work

In this thesis, we have investigated the problem of improving the exploration of the conformational space of a protein. The framework we use for protein conformational sampling is based on robotics-inspired expansive path planning algorithms. These algorithms use a low-dimensional projection to guide their search in the high-dimensional conformational space. Although the definition of this projection is essential to ensure good performance of the planning algorithms, little work had been devoted to this problem.

The contribution of the first part of this thesis consists of proposing a methodology to define "good" projections that accelerate the conformational search. Using the biological knowledge available for a given protein, it is possible to define a so-called "expert" projection that can efficiently guide the search through the high-dimensional conformational space of this protein. We evaluated the use of such expert projections for four medium-sized proteins. We demonstrated that our expert projections perform consistently better than randomly-defined or poorly-defined (so-called "misguided") projections. Our results show that using the expert projection increases the success rate of the planning algorithm at finding a transition pathway between two conformations, and improves computational runtime. Furthermore, using an expert projection

allows the planning algorithm to produce a better coverage of the projection space.

In the second part of this thesis, we proposed and evaluated the algorithm for constructing a reasonable projection automatically. The developed algorithm incorporates the defined concepts of the expert projections: it produces the sparse matrix with non-zero elements corresponding to the most flexible parts of a protein. The decomposition of a protein into fragments corresponding to its flexible regions has been already implemented in the SIMS framework (see Section 3.2) previously, and can be performed automatically (based on the secondary structure of a protein) or by an expert. The fragments represent the parts of a protein which are the most flexible and most likely to be involved in a transition. As such, they naturally define the regions that should be included in a "good" guiding projection. The produced auto projections demonstrate a consistent improvement of the conformational exploration for the directed search problem: for all considered proteins the auto projections produce higher probability of discovering the start-goal pathway in the given time than the random projections do. Due to the limitations of the chosen heuristics for identifying the most important protein parts the auto projections do not always yield a vast exploration of the conformational space. As a part of our future work, we will continue enhancing the auto-generated projections by experimenting with various methods that could automatically produce additional knowledge about studied proteins, such as Normal Mode Analysis (NMA), Principal Component Analysis (PCA), or graph-theory-based rigidity analysis among others. We are also planning to analyze

the influence of the dimensionality of the projections.

In the final part of this thesis we put the state-of-the-art technique of using linear projections to guide the conformational exploration under question. To assess the effect of projection on the conformational search we defined a special single-cell projection as a base case for comparison with all other projections. A single-cell projection maps the entire produced conformational data-structure into a single cell. Such projection eliminates the process of directing the exploration into a specific region of the conformational space by picking a projection cell to which this region corresponds. Thus, a single-cell projection does not introduce a projection bias, and its exploration might be considered as exploration without a projection. We compared the performance of two sampling-based planners leading the conformational exploration with and without a projection bias. The obtained results suggest that in some cases a *good* projection could enhance the conformational search, while in other cases even a good projection diminishes the performance of the planner and it is better not to use a projection at all. The discovered limitation of linear projections might be the outcome of the fact that protein motion has a highly non-linear nature. One of the possible directions for future work is the investigation of non-linear dimensionality reduction techniques in the context of conformational search problem.

One of the major contributions of the final part of the thesis consists of introducing the theory of how multi-cell linear projections operate that explains the presented results. In our future work we will continue to investigate this very important question

of the applicability of the linear projections as well as rigorously test the described theory of the projection effect. In light of the observations in the final part of this thesis, we are thinking of developing an alternative methods to guide the exploration of the high-dimensional space. As a part of our future work, we would like to build a planner that on each step would pick a parent conformation based on where it is located within the graph itself: a preference would be given to conformations in longer branches (a long branch has a higher probability of following a narrow passage) as well as to conformations that are closer to the leaves level (to increase the probability of exploring new regions).

Improving the process of conformational sampling in extremely high-dimensional spaces can open new horizons for studies of proteins by enabling modeling of larger proteins, such as viruses with several thousands of residues.

Bibliography

[1] H. Carlson, "Protein flexibility is an important component of structure-based drug discovery," *Curr. Pharm. Design*, vol. 8, no. 17, pp. 1571–1578, 2002.

[2] S. Adcock and J. McCammon, "Molecular dynamics: survey of methods for simulating the activity of proteins," *Chem. Rev.*, vol. 106, no. 5, pp. 1589–1615, 2006.

[3] D. Case, "Normal mode analysis of protein dynamics," *Curr. Opin. Struc. Biol.*, vol. 4, no. 2, pp. 285–290, 1994.

[4] E. Fuglebakk, N. Reuter, and K. Hinsen, "Evaluation of protein elastic network models based on an analysis of collective motions," *J. Chem. Theory Comput.*, vol. 9, no. 12, pp. 5618–5628, 2013.

[5] W. Liu, A. Ishchenko, and V. Cherezov, "Preparation of microcrystals in lipidic cubic phase for serial femtosecond crystallography," *Nat. Protocols*, vol. 9, pp. 2123–2134, 09 2014.

[6] J. Lipfert and S. Doniach, "Small-angle X-Ray scattering from RNA, proteins, and protein complexes," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 36, no. 1, pp. 307–327, 2007.

[7] C. Dominguez, M. Schubert, O. Duss, S. Ravindranathan, and F. H. Allain, "Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy," *Prog. Nucl. Magn. Reson. Spectrosc.*, vol. 58, no. 12, pp. 1–61, 2011.

[8] I. Al-Bluwi, T. Siméon, and J. Cortés, "Motion planning algorithms for molecular simulations: A survey," *Comput. Sci. Rev.*, vol. 6, no. 4, pp. 125–143, 2012.

[9] B. Gipson, D. Hsu, L. E. Kavraki, and J.-C. Latombe, "Computational models of proteins kinematics and dynamics: Beyond simulation," *Annu. Rev. Anal. Chem.*, vol. 5, no. 1, pp. 273–291, 2012.

[10] D. Hsu, J.-C. Latombe, and R. Motwani, "Path Planning in Expansive Configuration Spaces," *Int. J. Comput. Geom. Ap.*, vol. 9, no. 4-5, pp. 495–512, 1999.

[11] I. A. Şucan and L. E. Kavraki, "Kinodynamic motion planning by Interior-Exterior Cell Exploration," in *Algorithmic Foundation of Robotics VIII* (G. Chirikjian, H. Choset, M. Morales, and T. Murphey, eds.), vol. 57, pp. 449–464, Springer Berlin Heidelberg, 2010.

[12] A. E. García, "Large-amplitude nonlinear motions in proteins," *Phys. Rev. Lett.*, vol. 68, pp. 2696–2699, 1992.

[13] T. D. Romo, J. B. Clarage, D. C. Sorensen, and G. N. Phillips, "Automatic identification of discrete substates in proteins: Singular value decomposition

analysis of time-averaged crystallographic refinements," *Proteins*, vol. 22, no. 4, pp. 311–321, 1995.

[14] L. Skjaerven, S. M. Hollup, and N. Reuter, "Normal mode analysis for proteins," *J. Mol. Struc-theochem.*, vol. 898, no. 1-3, pp. 42–48, 2009.

[15] M. L. Teodoro, G. N. Phillips Jr., and L. E. Kavraki, "Singular value decomposition of protein conformational motions: Application to HIV-1 protease," in *Currents in Computational Molecular Biology*, pp. 198–199, Universal Academy Press Inc., 2000.

[16] A. Amadei, A. Linssen, B. De Groot, and H. Berendsen, "Essential degrees of freedom of proteins," *Molecular Engineering*, vol. 5, no. 1-3, pp. 71–79, 1995.

[17] D. Devaurs, K. Molloy, M. Vaisset, A. Shehu, T. Siméon, and J. Cortés, "Characterizing energy landscapes of peptides using a combination of stochastic algorithms," *IEEE Trans. Nanobiosci.*, vol. 14, no. 5, pp. 545–552, 2015.

[18] S. Thomas, G. Song, and N. M. Amato, "Protein folding by motion planning," *Phys. Biol.*, vol. 2, no. 4, p. S148, 2005.

[19] J. Cortés, T. Siméon, M. Remaud-Siméon, and V. Tran, "Geometric algorithms for the conformational analysis of long protein loops," *J. Comput. Chem.*, vol. 25, no. 7, pp. 956–967, 2004.

[20] B. Raveh, A. Enosh, O. Schueler-Furman, and D. Halperin, "Rapid sampling of molecular motions with prior information constraints," *PLoS Comput. Biol.*, vol. 5, no. 2, p. e1000295, 2009.

[21] D. J. Jacobs and M. F. Thorpe, "Generic rigidity percolation: The pebble game," *Phys. Rev. Lett.*, vol. 75, pp. 4051–4054, 1995.

[22] S. Thomas, X. Tang, L. Tapia, and N. M. Amato, "Simulating protein motions with rigidity analysis," *J. Comput. Biol.*, vol. 14, no. 6, pp. 839–855, 2007.

[23] N. Fox, F. Jagodzinski, Y. Li, and I. Streinu, "Kinari-web: a server for protein rigidity analysis," *Nucleic Acids Research*, vol. 39, no. suppl 2, pp. W177–W183, 2011.

[24] D. Luo and N. Haspel, "Multi-resolution rigidity-based sampling of protein conformational paths," *BCB*, pp. 786–792, 2013.

[25] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an ab-initio tree-based exploration," *The International Journal of Robotics Research*, vol. 29, no. 8, pp. 1106–1127, 2010.

[26] B. Olson, K. Molloy, S. F. Hendi, and A. Shehu, "Guiding probabilistic search of the protein conformational space with structural profiles," *J. Bioinform Comput Biol.*, vol. 10, no. 03, p. 1242005, 2012.

[27] K. Molloy and A. Shehu, "Interleaving global and local search for protein motion computation," in *Bioinformatics Research and Applications* (R. Harrison, Y. Li, and I. Mndoiu, eds.), vol. 9096 of *Lecture Notes in Computer Science*, pp. 175–186, Springer International Publishing, 2015.

[28] P. J. Ballester and W. G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes," *J. Comput. Chem.*, vol. 28, no. 10, pp. 1711–1723, 2007.

[29] M. Porto, U. Bastolla, H. E. Roman, and M. Vendruscolo, "Reconstruction of protein structures from a vectorial representation," *Phys. Rev. Lett.*, vol. 92, p. 218101, May 2004.

[30] I. A. Şucan and L. E. Kavraki, "On the performance of random linear projections for sampling-based motion planning," in *IEEE/RSJ*, pp. 2434–2439, 2009.

[31] B. Gipson, M. Moll, and L. E. Kavraki, "SIMS: A hybrid method for rapid conformational analysis," *PLoS ONE*, vol. 8, no. 7, p. e68826, 2013.

[32] R. Das and D. Baker, "Macromolecular modeling with Rosetta," *Annu. Rev. Biochem.*, vol. 77, no. 1, pp. 363–382, 2008.

[33] K. W. Kaufmann, G. H. Lemmon, S. L. Deluca, J. H. Sheehan, and J. Meiler, "Practically useful: what the Rosetta protein modeling suite can do for you," *Biochemistry*, vol. 49, no. 14, pp. 2987–98, 2010.

[34] M. Levitt, "A simplified representation of protein conformations for rapid simulation of protein folding," *J. Mol. Biol.*, vol. 104, no. 1, pp. 59–107, 1976.

[35] I. A. Şucan, M. Moll, and L. E. Kavraki, "The Open Motion Planning Library," *IEEE Robotics & Automation Magazine*, 2012.

[36] J. L. William Johnson, "Extensions of lipschitz mappings into a hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[37] I. Botos, B. R. O'Keefe, S. R. Shenoy, L. K. Cartner, D. M. Ratner, P. H. Seeberger, M. R. Boyd, and A. Wlodawer, "Structures of the complexes of a potent anti-HIV protein Cyanovirin-N and high mannose oligosaccharides," *J. Biol. Chem.*, vol. 277, no. 37, pp. 34336–42, 2002.

[38] N. J. Anthis, M. Doucleff, and G. M. Clore, "Transient, sparsely populated compact states of Apo and calcium-loaded Calmodulin probed by paramagnetic relaxation enhancement: Interplay of conformational selection and induced fit," *J. Am. Chem. Soc.*, vol. 133, no. 46, pp. 18966–18974, 2011.

[39] P. C. Whitford, O. Miyashita, Y. Levy, and J. N. Onuchic, "Conformational transitions of adenylate kinase: switching by cracking," *J. Molec. Biol.*, vol. 366, no. 5, pp. 1661–1671, 2007.

[40] A. J. Björkman, R. A. Binnie, H. Zhang, L. B. Cole, M. A. Hermodson, and S. L. Mowbray, "Probing protein-protein interactions. The ribose-binding protein in

bacterial transport and chemotaxis," *J. Biol. Chem.*, vol. 269, no. 48, pp. 30206–11, 1994.

[41] A. J. Björkman and S. L. Mowbray, "Multiple open forms of ribose-binding protein trace the path of its conformational change," *J. Mol. Biol.*, vol. 279, no. 3, pp. 651–64, 1998.